IJCSIS Vol. 11 No. 10, October 2013
ISSN 1947-5500

International Journal of Computer Science & Information Security


Cornell University
Library


Cogprints


Google scholar


.docstoc
find and share professional documents


ScientificCommons

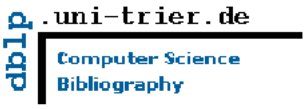

View my documents on
Scribd


BASE
Bielefeld Academic Search Engine


SCIRUS
search engine for science


SciRate.com


CiteSeerx beta


dblp .uni-trier.de
Computer Science
Bibliography


Q·Sensei


DOAJ
DIRECTORY OF
OPEN ACCESS
JOURNALS


EBSCO
HOST


ProQuest

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2013 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Cornell University Library, ScientificCommons, EBSCO, ProQuest and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Managing Editor

International Journal of Computer Science and Information Security *(IJCSIS – established since May 2009), is a venue to disseminate research and development results of lasting significance in the theory, design, implementation, analysis, and application of computing and security. As a scholarly open access peer-reviewed international journal, the primary objective is to provide the academic community and industry a forum for ideas and for the submission of original research related to Computer Science and Security. High caliber authors are solicited to contribute to this journal by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the Computer Science & Security.*

IJCSIS archives all publications in major academic/scientific databases; abstracting/indexing, editorial board and other important information are available online on homepage. Indexed by the following International agencies and institutions: Google Scholar, Bielefeld Academic Search Engine (BASE), CiteSeerX, SCIRUS, Cornell's University Library EI, Scopus, DBLP, DOI, ProQuest, EBSCO. Google Scholar reported a large amount of cited papers published in IJCSIS. IJCSIS supports the Open Access policy of distribution of published manuscripts, ensuring "free availability on the public Internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of [published] articles".

IJCSIS editorial board consisting of international experts solicits your contribution to the journal with your research papers, projects, surveying works and industrial experiences. IJCSIS is grateful for all the insights and advice from authors & reviewers.

We look forward to your collaboration. For further questions please do not hesitate to contact us at ***ijcsiseditor@gmail.com****.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 11, No. 10, October 2013 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

IJCSIS EDITORIAL BOARD

TABLE OF CONTENTS

Abstract — Vehicular ad hoc networks are networks in which no simultaneous end-to-end path exists. Typically, message delivery experiences long delays as a result of the disconnected nature of the network. In this form of network, our main goal is to deliver the messages to the destination with minimum delay. We propose relay assisted epidemic routing scheme in which we tend to use relay nodes (stationary nodes) at an intersection with a completely different number of mobile nodes which differs from existing routing protocols on how routing decision are made at road intersection where relay nodes are deployed. Vehicles keep moving and relay nodes are static. The purpose of deploy relay nodes is to increase the contact opportunities, reduce the delays and enhance the delivery rate. With various simulations it has been shown that relay nodes improves the message delivery probability rate and decreases the average delay.

Keywords - VANETs, Routing Protocols, Relay Nodes.

4. Paper 30091329: Blind Optimization for Data Warehouse during Design (pp. 27-31)

Rachid El mensouri and Omar El beqali, LIILAN/ GRM2I FSDM, Sidi Mohammed Ben Abdellah University, Fes, Morocco
Ziyati Elhoussaine, RITM laboratory ENSEM - ESTC - University Hassan II Ain chock, Casablanca, Morocco

Abstract — Design a suitable data warehouse is getting increasingly complex and requires more advance technique for different step. In this paper, we present a novel data driven approach for fragmentation based on the principal components analysis (PCA). Both techniques has been treated in many works [2][7]. The possibility of its use for horizontal and vertical fragmentation of data warehouses (DW), in order to reduce the time of query execution. We focus the correlation matrices, the impact of the eigenvalues evolution on the determination of suitable situations to achieve the PCA, and a study of criteria for extracting principal components. Then, we proceed to the projection of individuals on the first principal plane, and the 3D vector space generated by the first three principal components.
We try to determine graphically homogeneous groups of individuals and therefore, a horizontal fragmentation schema for the studied data table.

Keywords - data warehouse; optimization; PCA; vertical fragmentation; horizontal fragmentation; OLAP queries.

5. Paper 30091330: Traffic Intensity Estimation on Mobile Communication Using Adaptive Extended Kalman Filter (pp. 32-37)

Rajesh Kumar, Department of Electronics and Communication, Jabalpur Engineering College, Jabalpur, India (482011)
Agya Mishra, Department of Electronics and Communication, Jabalpur Engineering College, Jabalpur, India (482011)

Abstract — Traffic estimation is an important task in network management and it makes significant sense for network planning and optimization. The biggest challenge is to control congestion in the network and the blocking probability of call to provide users a barrier less communication. Effective capacity planning is necessary in controlling congestion and call drop rates in mobile communication thus an accurate prediction of traffic results congestion control, effective utilization of resources, network management etc. In this paper a real time mobile traffic data of different mobile service providers are estimated using adaptive Extended Kalman Filter method. This paper evaluates compares and concludes that, this approach is quite efficient with min normalized root mean square error (NRMSE) and it can be used for real time mobile traffic intensity estimation.

Keywords—Traffic Intensity Estimation, recursive filter, Mobile Traffic data, Extended Kalman filter, NRMSE.

6. Paper 30091334: Test Framework Development Based Upon DO-278A Using ISTQB Framework (pp. 38-41)

Raha Ashrafi, Computer Engineering Department, North Tehran Branch , Islamic Azad University, Tehran , Iran
Ramin Nassiri PhD, Computer Engineering Department, Central Tehran Branch, Islamic Azad University

Tehran, Iran

Abstract — DO-278A is an FAA Standard which was published for CNS/ATM software development in 2011 while ISTQB1 is the worldwide software testing standard framework. Software which are developed according to DO-289A are verified by Review, Analysis and a few other methods but they are not actually sufficient for testing so it would be quite reasonable to deploy ISTQB techniques for that purpose. This paper intends to show that ISTQB techniques may successfully work for DO-278A verification.

Keywords: component; Software Aviation system; CNS/ATM; DO-278A;, ISTQB; Verification.

7. Paper 30091344: Queuing Mechanism for WCDMA Network (pp. 42-47)

Vandana Khare, Associate professor & HOD ECE, RITW, Hyderabad, India
Dr. Y. Madhavee Latha, Professors & Principal MRECW, Secunderabad, India
Dr. D. SrinivasRao, Professors & Head ECE Department, JNTU, Hyderabad, India

Abstract - The network traffic in the upcoming wireless network is expected to be extremely nonstationary and next generation wireless networks including 3rd generation are expected to provide a wide range of multimedia services with different QoS constraints on mobile communication because of that there is no guarantee for a given system to provide good quality of service. So that there is a need to design an efficient queuing mechanism by which the QoS is going to be improved for wideband services. This paper proposes an efficient active queue management mechanism to control the congestion at the router. This paper is mainly aimed towards the WCDMA scheme for wideband services like video. So that it integrates the WCDMA with IR-RAN using active queue management. By simulation result our proposed WCDMA architecture along with active queue management (AQM) will achieves the effective peak signal to noise ratio (PSNR).

Keywords: WCDMA, IP-RAN, QoS, PSNR.

8. Paper 30091301: Implementation of Radial Basis Function and Functional Back Propagation Neural Networks for Estimating Inframe Fill Stability (pp. 48-53)

P. Karthikeyan, Research Scholar, Department of Civil Engineering, CMJ University.
Dr. S. Purushothaman, Professor, PET Engineering College, Vallioor-627117.

Abstract - ANSYS 14 software is used for analyzing the infill frames. The numerical values of strain are used to train the artificial neural network (ANN) topology by using Back propagation algorithm (BPA) and Radial basis function network (RBF). The training patterns used for the ANN algorithms are chosen from the strain data generated using ANSYS program. During the training process, node numbers are presented in the input layer of the ANN and correspondingly, strain values are presented in the output layer of the ANN. Depending upon the type of values present in the patterns, the learning capability of the ANN algorithms varies.

Keywords: Artificial Neural Network (ANN); Back propagation algorithm (BPA); Radial basis function (RBF).

9. Paper 30091302: Impact of Part-Of-Speech Based Feature Selection in Text Summarization (pp. 54-60)

Rajesh Wadhvani, Computer Science Department, National Institute Of Technology, Bhopal, India
R. K. Pateriya, Computer Science Department, National Institute Of Technology, Bhopal, India
Devshri Roy, Computer Science Department, National Institute Of Technology, Bhopal, India

Abstract — The standard practice in the field of summarization is to have a standard reference summary based on the queries. The summaries are manually generated by human experts. The automated summaries are then compared with the human generated summaries. In this paper a model which is useful for query-focused multi-document summarization is proposed. This model summarises documents of tagged data using cosine relevance measurement mechanism. Tagging is achieved by using the Stanford POS Tagger Package. We have used the concept of rarity in

addition to traditional raw term frequency for assigning the weights to the features. In this work for a given sentence out of all possible tag sequences best is derived by using argmax computation. For evaluation DUC-2007 dataset is used. The summaries generated by our technique are compared with the stranded summaries provided by DUC 2007 using ROUGE (Recall-Oriented Understudy for Gisting Evaluation). The performance metrics used for compassion are Recall, Precision, F-score.

Keywords: Text summarization, query-based summaries, sentence extraction.

10. Paper 30091304: Detection of Microcalcification in Breast Using Radial Basis Function (pp. 61-67)

Shaji B., Research Scholar, Vels University, Pallavaram, Chennai, India-600117.
Dr. Purushothaman S., Professor, PET Engineering College, Vallioor, INDIA-627117,
Rajeswari R., Research scholar, Mother Teresa Women's University, Kodaikanal-624102, INDIA.

Abstract - This paper presents combination of wavelet with radial basis function (RBF) in identifying the microcalcification (MC) in a mammogram image. Mammogram image is decomposed using Coiflet wavelet to 5 levels. Statistical features are extracted from the wavelet coefficients. These features are used as inputs to the RBF neural network along with a labeling of presence or absence of MC. The classification performance of RBF is minimum 95% out of the presence of total MC in a given mammogram.

Keywords: mammogram, Microcalcification, Radial basis function, Coiflet wavelet

11. Paper 30091307: Implementation of Intrusion Detection using BPARBF Neural Networks (pp. 68-72)

Kalpana Y., Research Scholar, VELS University, Pallavaram, Chennai, India-600117
Dr.Purushothaman S.,Professor, PET Engineering College, Vallioor, India-627117,
Rajeswari R., Research scholar, Mother Teresa Women's University, Kodaikanal-624102, India.

Abstract - Intrusion detection is one of core technologies of computer security. It is required to protect the security of computer network systems. Due to the expansion of high-speed Internet access, the need for secure and reliable networks has become more critical. The sophistication of network attacks, as well as their severity, has also increased recently. This paper focuses on two classification types: a single class (normal, or attack), and a multi class (normal, DoS, PRB, R2L, U2R), where the category of attack is detected by the combination of Back Propagation neural network (BPA) and radial basis function (RBF) Neural Networks. Most of existing IDs use all features in the network packet to look for known intrusive patterns. A well-defined feature extraction algorithm makes the classification process more effective and efficient. The Feature extraction step aims at representing patterns in a feature space where the attack patterns are attained. In this paper, a combination of BPA neural network along with RBF networks are used s for detecting intrusions. Tests are done on KDD-99 data set.

Keywords: network intrusion detection, kdd-99 datasets, BPARABF neural networks

12. Paper 30091308: Implementation of Daubauchi Wavelet with Radial Basis Function and Fuzzy Logic in Identifying Fingerprints (pp. 73-78)

Guhan P., Research Scholar, Department of MCA, VELS University, Chennai–600 117, India
Dr. Purushothaman S. ,Professor, PET Engineering College, Vallioor, India-627117,
Rajeswari R., Research scholar, Mother Teresa Women's University, Kodaikanal-624102, India.

Abstract - This paper implements wavelet decomposition for extracting features of fingerprint images. These features are used to train the radial basis function neural network and Fuzzy logic for identifying fingerprints. Sample finger prints are taken from data base from the internet resource. The fingerprints are decomposed using daubauchi wavelet 1(db1) to 5 levels. The coefficients of approximation at the fifth level is used for calculating

statistical features. These statistical features are used for training the RBF network and fuzzy logic. The performance comparisons of RBF and fuzzy logic are presented.

Keywords- Fingerprint; Daubauchi wavelet, radial basis function, fuzzy logic.

13. Paper 30091309: Implementation of Hidden Markov Model and Counter Propagation Neural Network for Identifying Human Walking Action (pp. 79-85)

Sripriya P., Research Scholar, Department of MCA, VELS University, Pallavaram, Chennai, India-600117,
Dr. Purushothaman S., Professor,PET Engineering College, Vallioor, India-627117,
Rajeswari R., Research scholar, Mother Teresa Women's University, Kodaikanal, India-624101.

Abstract - This paper presents the combined implementation of counter propagation network (CPN) along with hidden Markov model (HMM) for human activity recognition. Many methods are in use. However, there is increase in unwanted human activity in the public to achieve gainsay without any hard work. Surveillance cameras are installed in the crowded area in major metropolitan cities in various countries. Sophisticated algorithms are required to identify human walking style to monitor any unwanted behavior that would lead to suspicion. This paper presents the importance of CPN to identify the human GAIT.

Keywords: GAIT; human walking action; counter propagation network; hidden Markov model.

14. Paper 30091313: A Brief Study of Data Compression Algorithms (pp. 86-94)

Yogesh Rathore, CSE,UIT, RGPV, Bhopal, M.P., India,
Manish k. Ahirwar, CSE,UIT, RGPV, Bhopal, M.P., India
Rajeev Pandey, CSE,UIT, RGPV, Bhopal, M.P., India

Abstract — This paper present survey of several lossless data compression techniques and its corresponding algorithms. A set of selected algorithms are studied and examined. This paper concluded by stating which algorithm performs well for text data.

Keywords - Compression; Encoding; REL; RLL; Huffman; LZ; LZW;

15. Paper 30091314: fMRI image Segmentation using conventional methods versus Contextual Clustering (pp. 95-98)

Suganthi D., Research Scholar, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India-624101,
Dr. Purushothaman S, Professor, PET Engineering College, Vallioor, India-627117.

Abstract - Image segmentation plays a vital role in medical imaging applications. Many image segmentation methods have been proposed for the process of successive image analysis tasks in the last decades. The paper has considered fMRI segmentation in spite of existing techniques to segment the fMRI slices. In this paper an fmri image segmentation using contextual clustering method is presented. Matlab software 'regionprops' function has been used as one of the criteria to show performance of CC. The CC segmentation shows more segmented objects with least discontinuity within the objects in the fMRI image. From the experimental results, it has been found that, the Contextual clustering method shows a better segmentation when compared to other conventional segmentation methods.

Keywords: *Contextual clustering; segmentation; fMRI image.*

16. Paper 30091318: Implementation of Human Tracking using Back Propagation Algorithm (pp. 99-102)

Pratheepa S., Research Scholar, Mother Teresa Women's University, Kodaikanal, India-624101.
Dr. S. Purushothaman, Professor, PET Engineering College, Vallioor, India-627117.

Abstract - Identifying moving objects from a video sequence is a fundamental and critical task in many computer-vision applications. A common approach is to perform background subtraction, which identifies moving objects from the portion of a video frame that differs significantly from a background model. In this work, a new moving object-tracking method is proposed. The moving object is recorded in video. The segmentation of the video is done. Two important properties are used to process the features of the segmented image for highlighting the presence of the human. An artificial neural network with supervised back propagation algorithm learns and provides a better estimate of the movement of the human in the video frame. A multi target human tracking is attempted.

Keywords- Back propagation algorithm (BPA), Human tracking, and Video segmentation

17. Paper 30091331: Stock Market Trend Analysis Using Hidden Markov Models (pp. 103-110)

Kavitha G., School of Applied Sciences, Hindustan University, Chennai, India.
Udhayakumar A., School of Computing Sciences, Hindustan University, Chennai, India
Nagarajan D., Department of Information Technology, Salalah College of Technology, Salalah, Sultanate of Oman

Abstract — Price movements of stock market are not totally random. In fact, what drives the financial market and what pattern financial time series follows have long been the interest that attracts economists, mathematicians and most recently computer scientists [17]. This paper gives an idea about the trend analysis of stock market behaviour using Hidden Markov Model (HMM). The trend once followed over a particular period will sure repeat in future. The one day difference in close value of stocks for a certain period is found and its corresponding steady state probability distribution values are determined. The pattern of the stock market behaviour is then decided based on these probability values for a particular time. The goal is to figure out the hidden state sequence given the observation sequence so that the trend can be analyzed using the steady state probability distribution (p) values. Six optimal hidden state sequences are generated and compared. The one day difference in close value when considered is found to give the best optimum state sequence.

Keywords-Hidden Markov Model; Stock market trend; Transition Probability Matrix; Emission Probability Matrix; Steady State Probability distribution

18. Paper 30091342: Computational Impact of Hydrophobicity in Protein Stability (pp. 111-116)

Geetika S. Pandey, Research Scholar, CSE dept., RGPV, Bhopal (M.P), India
Dr. R.C Jain, Director, SATI(D), Vidisha(M.P), India

Abstract - Among the various features of amino acids, the hydrophobic property has most visible impact on stability of a sequence folding. This is mentioned in many protein folding related work, in this paper we more elaborately discuss the computational impact of the well defined ‘hydrophobic aspect in determining stability’, approach with the help of a developed ‘free energy computing algorithm’ covering various aspects - preprocessing of an amino acid sequence, generating the folding and calculating free energy. Later discussing its use in protein structure related research work.

Keywords- amino acids, hydrophobicity, free energy, protein stability.

19. Paper 31081343: Survey On MAC Protocols Used In Co-Operation Network Using Relay Nodes (pp. 117-120)

Shalini Sharma, PG-Scholar, DoEC, SSSIST
Mukesh Tiwari, Associate Professor, DoEC, SSSIST

Jaikaran Singh, Associate Professor, DoEC, SSSTTS

Abstract — In this paper a survey on a relay based media access scheme has been proposed. It has been observed that any cooperative scheme gives better performance by availability of additional path using the concept of a relay nodes. Relay based schemes more than one relay nodes are selected to improve the performance, so that if one fails the other can be used as a back. Such a co-operative scheme will enhance the performance of the network.

Keywords: MAC, cooperation, Relay Performance, Newtwork. Scheme.

The Application of Data Mining to Build Classification Model for Predicting Graduate Employment

Bangsuk Jantawan*
Department of Tropical Agriculture and International Cooperation
National Pingtung University of Science and Technology
Pingtung, Taiwan .

Cheng-Fa Tsai
Department of Management Information Systems
National Pingtung University of Science and Technology
Pingtung, Taiwan .

***Abstract*—Data mining has been applied in various areas because of its ability to rapidly analyze vast amounts of data. This study is to build the Graduates Employment Model using classification task in data mining, and to compare several of data-mining approaches such as Bayesian method and the Tree method. The Bayesian method includes 5 algorithms, including AODE, BayesNet, HNB, NaviveBayes, WAODE. The Tree method includes 5 algorithms, including BFTree, NBTree, REPTree, ID3, C4.5. The experiment uses a classification task in WEKA, and we compare the results of each algorithm, where several classification models were generated. To validate the generated model, the experiments were conducted using real data collected from graduate profile at the Maejo University in Thailand. The model is intended to be used for predicting whether a graduate was employed, unemployed, or in an undetermined situation.**

Keywords-Bayesian method; Classification model; Data mining; Tree method

I. INTRODUCTION

Graduates employability remains as national issues due to the increasing number of Graduates produced by higher education institutions each year. According to the United Nations Educational Scientific and Cultural Organization report, enrollment in higher education more than doubled over the past two decades from 68 million in 1991 to 151 million in 2008. At the same time, the financial crisis that began in 2008 has resulted in increasing unemployment, as highlighted in International Labor Organization's Global Employment Trends reports. The global unemployment rate was 6.2 percent in 2010 compared to 5.6 percent in 2007. According to the 2012 report, young people continue to be the hardest hit by the job crisis with 74.8 million youth being unemployed in 2011, an increase of more than 4 million since 2007 [1].

With many economies being reported as not generating sufficient employment opportunities to absorb growth in the working-age population, a generation of young productive workers will face an uncertain future unless something is done to reverse this trend. To increase the graduates' chances of obtaining decent jobs that match their education and training, universities need to equip their students with the necessary competencies to enter the labor market and to enhance their capacities to meet specific workplace demands [1].

As Thailand, there were 320,815 graduates in 2006 with bachelors' degrees and above. This figure increased to 371,982 in 2007, about 75.02 percent of graduates in 2006 (excluding those from open universities) were employed. About 18 percent of graduates were unemployed. The proportion of employed graduates dropped to 68.65 percent in 2008 and unemployment rose to 28.98 percent [2]. Hence, preparing young people to enter the labor market has therefore become a critical responsibility for universities [1].

According to data mining is a technology used to describe knowledge discovery and to search for significant relationships such as patterns, association and changes among variables in databases [3]. There are several of data mining techniques that can be used to extract relevant and interesting knowledge from large data. Data mining has several tasks such as classification and prediction, association rule mining and clustering. Moreover, classification is one of the most useful techniques in data mining to build classification models from an input data set. The used classification techniques commonly build models that are used to predict future data trends. There are several algorithms for data classification which include decision tree and Naïve Bayes classifiers and so on [4].

Furthermore, decision tree is one of the most used techniques, due to it creates the decision tree from the data given using simple equations depending mainly on calculation of the gain ratio, which gives automatically some sort of weights to attributes used, and the researcher can implicitly recognize the most effective attributes on the predicted target. As a result of this technique, a decision tree would be built with classification rules generated from it [5], and another classification that is Naïve Bayes classifier. This classification is used to predict a target class. It depends on calculations of probabilities, namely Bayesian theorem. Because of this use, results from the classifier are more accurate and more efficiency as well as more sensitive to new data added to the dataset [5].

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Therefore, the aim of this research is to predicting of graduate employment has been employed, unemployed or others within the first twelve months after graduation, the raw data received from the Planning Division Office of Maejo University in Thailand. With experiment realized through a data classification that classifies a graduate profile as employed, unemployed or others. Subsequently, the main contribution of this research is the comparison of classification accuracy between two algorithms from commonly used data mining techniques in the education domain in Waikato Environment for Knowledge Analysis (WEKA) environment.

II. LITERATURE REVIEW

Several researches used data mining techniques for extracting rules and predicting certain behaviors in several areas. For example, researcher has defined the performance of a frontline employee, as his/her productivity comparing with his/her peers [6]. On the other hand, described the performance of university teachers included in his study, as the number of researches cited or published. In general, performance is usually measured by the units produced by the employee in his/her job within the given period of time [7].

Researchers like Chein and Chen [8] have worked on the improvement of employee selection, by building a model, using data mining techniques, to predict the performance of newly applicants. Depending on attributes selected from their curriculum vitae, job applications and interviews. Their performance could be predicted to be a base for decision makers to take their decisions about either employing these applicants or not. And they also used several attributes to predict the employee performance. They specified gender, age, experience, marital status, education, major subjects and school tires as potential factors that might affect the performance. Then they excluded age, gender and marital status, so that no discrimination would exist in the process of personal selection. As a result for their study, they found that employee performance is highly affected by education degree, the school tire, and the job experience.

Moreover, researchers also are identified three major requirements concerned by the employers in hiring employees, which are basic academic skills, higher order thinking skills, and personal qualities. The work is restricted in the education domain specifically analyzing the effectiveness of a subject, English for Occupational Purposes in enhancing employability skills [9], [10]. Subsequently, Kahya [11] also searched on certain factors that affect the job performance. The researcher reviewed previous studies, describing the effect of experience, salary, education, working conditions and job satisfaction on the performance. As a result of the research, it has been found that several factors affected the employee's performance. The position or grade of the employee in the company was of high positive effect on his/her performance. Working conditions and environment, on the other hand, had shown both positive and negative relationship on performance. Highly educated and qualified employees showed dissatisfaction of bad working conditions and thus affected their performance negatively. Employees of low qualifications, on the other hand, showed high performance in spite of the bad conditions. In addition, experience showed positive relationship in most cases, while education did not yield clear relationship with the performance.

More recently, in Malaysia proposes a new Malaysian Engineering Employability Skills Framework, which is constructed based on requirement by accrediting bodies and professional bodies and existing research findings in employability skills as a guideline in training package and qualification of country. Nonetheless, not surprisingly, graduates employability is rarely being studied especially within the scope of data mining, mainly due to limited and authentic data source available [12].

Employability issues have also been taken into consideration in other countries. Research by the Higher Education Academy with the Council for Industry and Higher Education in United Kingdom concluded that there are six competencies that employers observe in individual who can transform the organizations and add values in their careers [13]. The six competencies are cognitive skills or brainpower, generic competencies, personal capabilities, technical ability, business or organization awareness and practical elements. Furthermore, it covers a set of achievements comprises skills, understandings and personal attributes that make graduates more likely to gain employment and successful in their chosen occupations which benefits the graduates, the community and also the economy.

However, data mining techniques have indeed been employed in education domain, for instance in prediction and classification of student academic performance using Artificial Neural Network [14], [15] and a combination of clustering and decision tree classification techniques [14]. Experiments in [16] classifies students to predict their final grade using six common classifiers (Quadratic Bayesian classifier, 1-nearest neighbour (1-NN), k-nearest neighbor (k-NN), Parzen-window, multilayer perceptron (MLP), and Decision Tree). With regards to student performance, researchers have discovered individual student characteristics that are associated with their success according to grade point averages (GPA) by using a Microsoft Decision Trees classification technique [17]. In addition, Kumar and Chadha [18] have shown some applications of data mining in educational institution that extracts useful information from the huge data sets. Data mining through analytical tool offers user to view and use current information for decision making process such as organization of syllabus, predicting the registration of students in an educational program, predicting student performance, detecting cheating in online examination as well as identifying abnormal/erroneous values.

Accordingly, the study showed a positive relationship between affiliation motivation and job performance in Malaysia. They have tested the influence of motivation on job performance for state government employees of country. As people with higher affiliation motivation and strong interpersonal relationships with colleagues and managers tend to perform much better in their jobs [19].

Tair and El-Halees [20] used data mining to improve graduate student's performance, and overcome the problem of low grades of graduate students using association, classification, clustering and outlier detection rules. Similar to

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the study of Bhardwaj and Pal [21] in which a data model was used to predict student's performance with emphasis on identifying the difference of high learners and slow learners using byes classification.

Decision tree as a classification algorithm has been utilized in [22] to predict the final grade of a student in a particular course. The same algorithm has been applied in Yadav, Bharadwaj, and Pal [23] on past student performance data to generate a model to predict student performance with highlights on identifying dropouts and students who need special attention and allow teachers to provide appropriate advising or counseling. Conversely, Pandey and Pal [24] have considered the qualities the teacher must possess in order to determine how to tackle the problems arising in teaching, key points to be remembered while teaching and the amount of knowledge of the teaching process. In the course of identifying significant recommendations, John Dewey's principle of bipolar and Reybem's tri-polar education systems have been used to establish a model to evaluate the teacher ship on the basis of student feedback using data mining. Consequently, While [25] also used classification technique to build models to predict new applicant's performance, they used the same to forecast employee's talents [26], [27], [28], [29].

Another technique called fuzzy has been applied in [30] build a practical model for improving the efficiency and effectiveness of human resource management while [31] has improved and employed it to evaluate the performance of employees of commercial banks.

Generally, this paper is a preliminary attempt to use data mining concepts, particularly classification, to help supporting the human resources directors and decision makers by evaluating employees' data to study the main attributes that may affect the employees' performance. The paper applied the data mining concepts to develop a model for supporting the prediction of the employees' performance. In section 2, a complete description of the study is presented, specifying the methodology, the results, discussion of the results. Among the related work, we found that work done by [11] is most related to this research, whereby the work mines historical data of students' academic results using different classifiers (Bayes, trees, function) to rank influencing factors that contribute in predicting student academic performance.

III. METHODOLOGY

The major objective of the proposed methodology is to build the classification model that classify a graduate profile as employed, unemployed or undetermined using data sourced from the Maejo University in Thailand for 3 academic years, which consists of 11,853 instances. To build the classifiers, we combines the Cross Industry Standard Process for Data Mining methodology [32] and Process of Knowledge Discovery in [33] in which data mining is a significant step. The iterative and sequence of steps are shown in figure 1. It consists of five steps, which include Business understanding, data understanding, data preparation, modeling, evaluation and deployment along with the data discovery processes such as data cleaning, data integration, data selection, data transformation, data mining, evaluation and presentation.

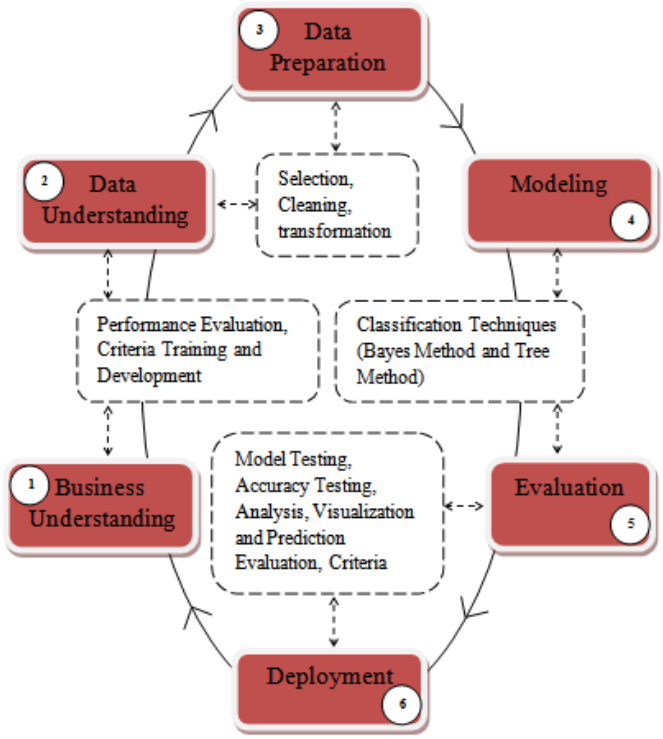


Figure 1. Framework of KDD.

A. Business Understanding or Data Classification Preliminaries

There are two-step processes of data classification. First step is training data, which called the learning step; a model that describes a predetermined set of classes or concepts is built by analyzing a set of training database instances. Each instance is assumed to belong to a predefined class. The second step is testing data; the model is tested using a different data set that is used to estimate the classification accuracy of the model. If the accuracy of the model is considered acceptable, the model can be used to classify future data instances for which the class label is not known. Finally, the model acts as a classifier in the decision making process. There are several techniques that can be used for classification such as decision tree, Bayesian methods and so on.

Decision tree classifiers are quite popular techniques because the construction of tree does not require any domain expert knowledge or parameter setting, and is appropriate for exploratory knowledge discovery. Decision tree can produce a model with rules that are human-readable and interpretable. Decision Tree has the advantages of easy interpretation and understanding for decision makers to compare with their domain knowledge for validation and justify their decision. Some of decision tree classifiers are C4.5/C5.0/J4.8, NBTree, and others [4].

The C4.5 technique is one of the decision tree families that can produce both decision tree and rule-sets; and construct a tree for the purpose of improving prediction accuracy. The C4.5/C5.0/J48 classifier is among the most popular and powerful decision tree classifiers. C4.5 creates an initial tree

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

using the divide-and-conquer algorithm. The full description of the algorithm can be found in any data mining or machine learning books such as Han and Kamber [33].

Furthermore, the WEKA was used for data mining. WEKA was developed at the University of Waikato in New Zealand [34]. It contains a large collection of state-of-the-art machine learning and data mining algorithms written in Java. WEKA is approved widely and is one of most complete tools in data mining. As a public data mining platform, WEKA gathers many machine learning algorithms to mine data, including data pretreatment, classification, regression, cluster class, association rule mining and visualization on new interface. WEKA has become very popular with academic and industrial researchers, and is also widely used for teaching purposes. WEKA toolkit package has its own version known as J48. J48 is an optimized implementation of C4.5.

B. Data Understanding

In order to classify a graduate profile as employed, unemployed or undetermined using data sourced from the Maejo University in Thailand for 1 academic years, the total are 5,361 instances in 2011. Table 1 shows the complete attributes for Graduate profile data source.

C. Data-Preprocessing

The raw data received from the Planning Division Office, Maejo University at Chiang Mai Province in Northern Thailand, which required the Data Pre-processing to prepare the dataset for the classification task. First, the data source has been transferred to Excel sheets and replaced them with the mean values of the attribute. Then, cleaning data involves eliminating data with missing values in critical attributes, correcting inconsistent data, identifying outliers, as well as removing duplicate data. For example, some attributes like GPA, have been entered in continuous values. Data source from the total of 5,361 instances in the raw data, the data cleaning process ended up 3,530 instances that are ready to be mined. These files are prepared and converted to (.csv) format to be compatible with the WEKA data mining is used in building the model.

D. Modeling

The section of modeling and experiments, the classification models have been built to predict the employment status (employed, unemployed, others) for graduate profiles. Using the decision tree technique, in this technique, the gain ratio measure is used to indicate the weight of effective of each attribute on the tested class, and accordingly the ordering of tree nodes is specified. The results are discussed in the following sections.

TABLE I. THE ATTRIBUTES OF THE GRADUATES EMPLOYMENT DATA AFTER THE PRE-PROCESSING

No.	Attributes	Values	Descriptions
1	Prefix	{Male, Female, Dr., Associate Prof. Dr,..}	Prefix of graduate
2	Gender	{ Male, Female}	Gender of the graduate (Male, Female)
3	Province	{Bangkok, Suratthani…}	Province of graduate
4	Degree	{Bachelor, Ph.D., Master}	Degree of graduate
5	Educational background	{B.Sc, B.L.A, M.A., M.B.A.}	Graduate background
6	Faculty	{Science, Agricultural Production, Economics,…}	Faculty of graduate
7	program	{Computer science, Information technology,}	Program of graduate
8	GPA	{Interval value}	GPA for current qualification
9	WorkProvince	{Bangkok, Suratthani…}	Province of Student's work
10	Status	{Employed, UnemployedandNotStudy,Study, …}	Work status of graduate
11	Talent	{Computer, Art, Food physical, …}	Talent of graduate
12	Position	{Chef, Trad, boss,...}	Position of graduate
13	Satisfaction	{ Pleased, Lack_of_consistence,Other,..}	Satisfaction of graduate with work
14	PeriodTimeFindwork	{FourToSix, OneToThree, SevenToNine,..}	Time of find work
15	WorkDirectGraduate	{Direct,NotDirect,NoIdentify}	Matching of Graduate education with graduate work
16	ApplyKnowlageWithWork	{Moderate, NoIdentify, Much,…}	Knowledge of graduate can apply with work
17	ResonNotWork	{Soldier, Business, NotFindWork…}	The Reason that don't have work of graduate
18	ProblemOfWork	{Lack_Of_support, NoProblem…}	The problem of work
19	RequirementsOfStudy	{NoNeed, Need}	Requirements of graduate to continue to study
20	LevelOfStudyRequired	{Master, Graduate_Diploma, NoIdentify,…}	Level Required to study of Graduate
21	InstitutionNeed	{Private, Aboard, Government…}	Institution Requirement to study of graduate

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The classification model is performed in two steps, which include training and testing. Once the classifier is constructed, testing dataset is used to estimate the predictive accuracy of the classifier. Then the WEKA have 4 types of testing option, which are using the training set, supplied test set, cross validation and percentage split. If we use training set as the test option, the test data will be sourced from the same training data, hence this will decrease reliable estimate of the true error rate. In the part of Supplied test set permit us to set the test data which been prepared separately from the training data. Cross-validation is suitable for limited dataset whereby the number of fold can be determined by user. 10-fold cross validation is widely used to get the best estimate of error. It has been proven by extensive test on numerous datasets with different learning techniques.

IV. Results and Discussion

Ten classification techniques Method have been applied the dataset to build the classification model. The techniques are: The decision tree with 5 versions, BFTree, NBTree, REPTree, ID3, C4.5 (J4.8 in WEKA), and Naïve Bayes classifier with 5 version such as the Averaged One-Dependence Estimators (AODE), BayesNet, HNB, NaiveBayes, the Weightily Averaged One-Dependence Estimators (WAODE).

In table 2 shows the Classification accuracy using various algorithms under Tree method in WEKA. In addition, the table provides comparative results for the kappa statistics, mean absolute error, root mean squared error, relative absolute error, and root relative squared error from the total of 1,059 testing instances. Subsequently, result of the J48 algorithm achieved the highest accuracy percentage as compared to other algorithms. The second accuracy is REPTree algorithm, BFTree, NBTree, ID3 Respectivel.

Furthermore, figure 2 also shows an example of tree structures of J48 algorithms. Graf adds nodes to the decision trees to increase predictive accuracy. Accordingly, table 3 shows the classification accuracies for various algorithms under Bayes method. The table provides comparative results for the kappa statistics mean absolute error, root mean squared error, relative absolute error, and root relative squared error from the total of 1,059 testing instances. Also, table 3 presents the WAODE algorithm achieved the highest accuracy percentage as compared to other algorithms. Despite treating each tree augmented naive Bayes equally, have extended the AODE by assigning weight for each tree augmented naive Bayes differently as the facts that each attributes do not play the same role in classification.

In addition, a performance comparison of the Bayesian and Tree methods shows that the WAODE algorithm achieved the highest accuracy of 99.77% using the graduate data set. The second highest accuracy was achieved using a tree method, the J48 algorithm, with an accuracy of 98.31%. Using the Bayes method, the AODE algorithm, was third, with a prediction accuracy of 98.30%. We found that both classification approaches were complementary because the Bayesian methods provide a better view of association or dependencies among attributes, whereas the results of the tree method are easier to interpret.

TABLE II. The classification accuracy using various algorithms under tree method in WEKA

Algorithm	Accuracy (%)	Error Rate (%)	Kappa Statistics	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
ID3	90.20	9.3229	0.9408	0.0126	0.112	6.37	35.56
J48	98.31	1.69	0.9586	0.0166	0.0912	7.886	28.0945
BFTree	98.24	1.76	0.9572	0.0165	0.0928	7.8327	28.5893
NBTree	98.07	1.93	0.953	0.01	0.0942	4.7667	29.0166
REPTree	98.27	1.73	0.9578	0.0169	0.0919	8.0104	28.3152

TABLE III. The classification accuracy using various algorithms under Bayes method in WEKA

Algorithm	Accuracy (%)	Error Rate (%)	Kappa Statistics	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
AODE	98.30	1.70	0.9586	0.0096	0.0909	4.5744	28.0238
BayesNet	98.02	1.98	0.9525	0.0124	0.0911	5.8937	28.0643
HNB	97.25	2.75	0.9331	0.0158	0.1097	0.1097	33.8206
NaiveBayes	97.96	2.04	0.9509	0.0118	0.0937	5.607	28.8735
WAODE	99.77	0.23	0.9946	0.0028	0.0324	1.3306	9.9742

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

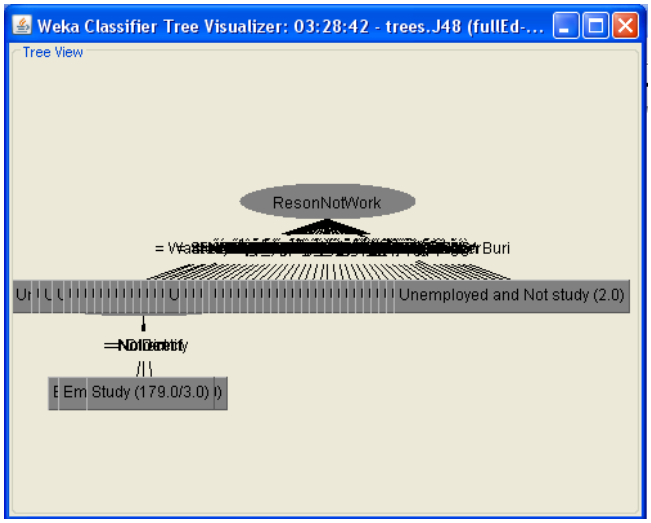


Figure 2. The tree structure for J48 algorithms.

Figure 3 shows the mapping of the root mean squared error values resulting from the classification experiment. This knowledge can be used to gain insights into the employment trend of graduates from local institutions of higher learning. A radial display of the root mean squared error across all algorithms under both Bayesian and tree-based methods reveals the accuracy of these approaches. A smaller mean squared error results in a better forecast. Based on this figure, Bayesian methods produced a better forecast than the corresponding tree methods.

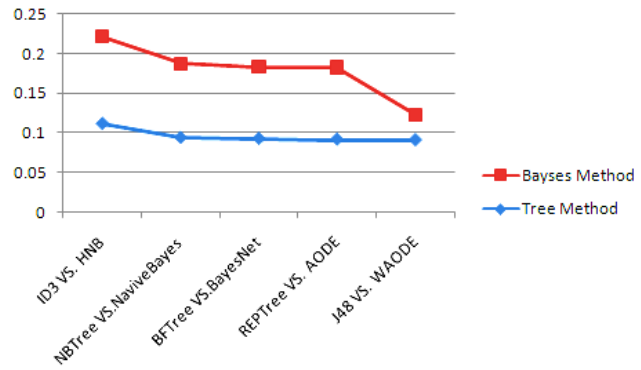


Figure 3. Mapping of the root mean squared error values of Bayesian and Tree methods.

V. Conclusions and Future Work

As graduates remains increase number of graduates produced by higher education institutions each year, graduates are facing more competition to ensure their employment in the job market. The purpose of the study is to assist higher-education institutions in equipping their graduates with sufficient skills to enter the job market. This study attempts to identify the attributes that influence graduate employment based on actual data obtained from the graduates themselves 12 months after graduation.

This study attempts to predict whether a graduate has been employed, remains unemployed, or is in an undetermined situation after graduation. We performed this prediction based on a series of classification experiments using various algorithms under Bayesian and decision methods to classify a graduate profile as employed, unemployed, or other. Results show that the WAODE algorithm, a variant of the Bayse algorithm, achieved the highest accuracy of 99.77%. The average accuracy of other Tree algorithms was 98.31%.

In future research, we hope to expand the data set from the tracer study to include more attributes and to annotate the attributes with information such as the correlation factor between current and previous employers. We are also looking at integrating data sets from different sources of data, such as graduate profiles from the alumni organizations of various educational institutions. We plan to introduce clustering in the preprocessing phase to cluster the attributes before attribute ranking. Finally, we may adopt other data-mining techniques, such as anomaly detection or classification-based association, to gain more knowledge of the graduate employability in Thailand. We also plan to use a data set from the National Pingtung University of Science and Technology in Taiwan and compare the results with the data set from Thailand.

Acknowledgment

B. Jantawan would like to express thanks Dr. Cheng-Fa Tsai, professor of the Management Information Systems Department, the Department of Tropical Agriculture and International Cooperation, National Pingtung University of Science and Technology in Taiwan for supporting the outstanding scholarship, and highly appreciates to Mr. Nara Phongphanich and the Planning Division Office, Maejo University in Thailand for giving the information.

References

[1] United Nations Educational Scientific and Cultural Organization, Graduate Employability in Asia. UNESCO Bangkok, Asia and Pacific Regional Bureau, Bangkok: Thailand, 2012.

[2] United Nations Educational Scientific and Cultural Organization, The Impact of Economic Crisis on Higher Education. UNESCO Bangkok, Asia and Pacific Regional Bureau, Bangkok: Thailand, 2012.

[3] C. Rygielski, J.C. Wang, D.C. Yen, "Data mining techniques for customer relationship management," Technology in Society, vol. 24, pp. 483–502, 2002.

[4] S.F. Shazmeen, M.M.A. Baig, and M.R. Pawar, "Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis," Journal of Computer Engineering, vol. 10(6), pp. 01-06, 2013.

[5] T. Fawcett, ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, Hewlett-Packard Company, Palo Alto: CA, 2003.

[6] O.M. Karatepe, O. Uludag, I. Menevis, L. Hadzimehmedagic, and L. Baddar, "The effects of selected individual characteristics on frontline employee performance and job satisfaction," Tourism Management, vol. 27, pp. 547–560, 2006.

[7] D. Schwab, "Contextual variables in employee performance-turnover relationships," Academy of Management Journal, vol. 34(4), pp. 966–975, 1991.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[8] C.F. Chein, L.F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high technology industry," Expert Systems with Applications, vol. 34, pp. 280–290, 2008.

[9] L.A. Shafie and S. Nayan, "Employability awareness among Malaysian undergraduates," International Journal of Business and Management, vol. 5(8), pp. 119–123, 2010.

[10] M. Mukhtar, Y. Yahya, S. Abdullah, A.R. Hamdan, N. Jailani, and Z. Abdullah, "Employability and service science: Facing the challenges via curriculum design and restructuring," In: International Conference on Electrical Engineering and Informatics, pp. 357–361, 2009.

[11] E. Kayha, "The Effects of job characteristics and working conditions on job performance," International Journal of Industrial Ergonomics, vol. 37, pp. 515–523, 2007.

[12] A. Zaharim, M.Z. Omar, Y.M. Yusoff, N. Muhamad, A. Mohamed, and R. Mustapha, "Practical framework of employability skills for engineering graduate in Malaysia," In: IEEE EDUCON Education Engineering 2010: The Future of Global Learning Engineering Education, pp. 921–927, 2010.

[13] C. Rees, P. Forbes, and B. Kubler, Student Employability Profiles: A Guide for Higher Education Practitioners, 2nd ed., The Higher Education Academy, York: United Kingdom, 2006.

[14] M. Wook, Y.H. Yahaya, N. Wahab, and M.R.M. Isa, "Predicting NDUM student's academic performance using Data mining techniques," In: Second International Conference on Computer and Electrical Engineering, pp. 357–361, 2009.

[15] E.N. Ogor, "Student academic performance monitoring and evaluation using Data mining techniques," In: Fourth Congress of Electronics, Robotics and Automotive Mechanics, pp. 354–359, 2007.

[16] B. Minaei-Bidgoli, D.A. Kashy, G. Kortemeyer, and W.F. Punch, "Predicting student performance: An application of Data mining methods with an educational web-based system," In: 33rd Frontiers in Education Conference, pp. 13–18, 2003.

[17] H. Guruler, A. Istanbullu, and M. Karahasan, "A new student performance analysing system using knowledge discovery in higher educational databases," Computers & Education. Vol. 55(1), pp 247–254, 2010.

[18] V. Kumar and A. Chadha, "An empirical study of the applications of Data mining techniques in higher education," International Journal of Advanced Computer Science and Applications, vol. 2(3), pp. 80–84, 2011.

[19] F. Salleh, Z. Dzulkifli, W.A. Abdullah, and N. Yaakob, "The effect of motivation on job performance of state government employees in Malaysia," International Journal of Humanities and Social Science, vol. 1(4), pp. 147–154, 2011.

[20] M.M.A. Tair and A.M. El-Halees, "Mining educational data to improve students' performance: A case study," International Journal of Information and Communication Technology Research, vol. 2(2), pp. 140–146, 2012.

[21] B.K. Bhardwaj and S. Pal, "Data mining: A prediction for performance improvement using classification," International Journal of Computer Science and Information Security, vol. 9(4), 2011.

[22] Q.A. Al-Radaideh, E.M. Al-Shawakfa, M.I. Al-Najjar, "Mining student data using decision trees," The International Arab Journal of Information Technology, 2006.

[23] S.K. Yadav, B. Bharadwaj, and S. Pal, "Data mining applications: A comparative study for predicting student's performance," International Journal of Innovative Technology and Creative Engineering, vol. 1(12), pp. 13–19, 2011.

[24] U.K. Pandey and S. Pal, "Mining data to find adept teachers in dealing with students," International Journal of Intelligent Systems and Applications, vol. 4(3), 2012.

[25] Q.A. Al-Radaideh and E.A. Nagi, "Using Data mining techniques to build a classification model for predicting employees performance," International Journal of Advanced Computer Science and Applications, vol. 3(2), pp. 144–15,12012.

[26] H. Jantan, A.R. Hamdan, and Z.A. Othman, "Knowledge Discovery techniques for talent forecasting in human resource application," International Journal of Human and Social Sciences, vol. 5(11), pp. 694–702, 2010.

[27] H. Jantan, A.R. Hamdan, and Z.A. Othman, "Human talent prediction in HRM using C4.5 classification algorithm," International Journal on Computer Science and Engineering. vol. 2(8), pp. 2526–2534, 2010.

[28] H. Jantan, A.R. Hamdan, and Z.A. Othman, "Towards applying data mining techniques for talent management," International Association of Computer Science & Information Technology.

[29] H. Jantan, A.R. Hamdan, and Z.A. Othman, "Talent knowledge acquisition using data mining classification techniques," 3rd Conference on Data mining and Optimization, 2011.

[30] H. Jing, "Application of fuzzy data mining algorithm in performance evaluation of human resource," Computer Science-Technology and Applications, vol. 1, 2009.

[31] H. Zhang, "Fuzzy evaluation on the performance of human resources management of commercial banks based on improved algorithm," Power Electronics and Intelligent Transportation System, vol. 1, 2009.

[32] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 2000.

[33] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann publishers, San Francisco: CA, 2006.

[34] I.H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann publishers, San Francisco: CA, 2005.

AUTHORS PROFILE

Bangsuk Jantawan is a Ph.D. student at the Department of Management Information System, National Pingtung University of Science and Technology (NPUST) in Taiwan, where her research is the Application of Data Mining to Build Classification Model for Predicting Graduate Employment. She obtained her MSc degree in Education Technology from the King Mongkut's University of Technology Thonburi, and a BSc degree in Computer Engineering form Dhurakij Pundit University in Bangkok, Thailand. Jantawan's prior research involved the Information System Development for Educational Quality Administration of Technical Colleges in Southern Regional Office of the Vocational Commission. Her present research interests include the data mining, education system development, decision making, and machine learning.

Cheng-Fa Tsai is full professor of the Management Information Systems Department at National Pingtung University of Science and Technology (NPUST), Pingtung, Taiwan. His research interests are in the areas of data mining and knowledge management, database systems, mobile communication and intelligent systems, with emphasis on efficient data analysis and rapid prototyping. He has published over 160 well-known journal papers and conference papers and several books in the above fields. He holds, or has applied for, nine U.S. patents and thirty ROC patents in his research areas.

Randomness Analysis of 128 bits Blowfish Block Cipher on ECB mode

[1]Ashwak ALabaichi [2]Ramlan Mahmod [3]Faudziah Ahmad

[1,3]Information Technology Department, Technology University Utara Malaysia
06010, Sintok, Malaysia
[1]Department of computer science, Faculty of Sciences Kerbala University, Kerbala,00964, Iraq

[2] Faculty of Computer Science and Information Technology University Putra Malaysia
Serdang, Selangor, Malaysia

Abstract

This paper presents the statistical test of randomness on the Blowfish Block Cipher 128-bit is continuation with our eaelier papers title" Randomness Analysis on Blowfish Block Cipher using ECB mode" and "Randomness Analysis on Blowfish Block Cipher using ECB and CBC Modes", . Blowfish128-bit is extention of blowfish 64-bit. Blowfish 128-bit resemble blowfish 64-bit but only in blowfish algorithm block size is 64 bits but in an extension version block size is 128 bits and all operations based on 64 bits instead of on 32 bits. Blowfish 128-bit is a symmetric block cipher with variable key lengths from 64 bits up to a maximum of 896 bits this leads to increase security of the algorithm. The randomness testing was performed using NIST Statistical Test Suite. The tests were performed on Blowfish 128-bit with ECB mode.Our analysis showed that Blowfish 128-bit algorithm with ECB mode such as blowfish 64-bit where is not inappropriate with text and images files that contain huge number of identical bytes and better than blowfish 64-bit with video files. c++ is used in the implementation of the blowfish 128-bit while NIST is implemented under Linux operating system.

Keywords: Block Cipher, Blowfish Algorithm 128-bit, ECB mode, randomness testing.

1. INTRODUCTION

Blowfish algorithm was designed by Schneier at the Cambridge Security Workshop in December 1993 to replace the Data Encryption Standard (DES). It has been widely analyzed and gradually accepted as a good and powerful encryption algorithm offering several advantages, among them its suitability and efficiency for implementing hardware. It is also unpatented and therefore does not require any license. The elementary operators of Blowfish algorithm comprise table lookup, addition and XOR with the table being made up of four S-boxes and a P-array. Based on Feistel rounds, Blowfish algorithm is a cipher with the F-function design being a simplified version of the principles employed in DES to provide similar security, faster speed and higher efficiency in software.

The effective cryptanalysis has not been present because its good encryption rate in software [1-3],[8]. Even though it is not as well-known as Advance Encryption Standard (AES), the uniqueness of Blowfish algorithm and the efficiency of its algorithm have led to its growing popularity in the open source community [4].

Meanwhile, ALabaichi, Mahmood, Ahmad in [17], and ALabaichi, Mahmood, Ahmad and Mechee in [18] uncovered issue with Blowfish. The issue lies in its compatibility with image and text files that involve large strings of identical bytes, in Particular, the problems related into randomness of the output with encrypted text and image files.

Nechvatal et al. in [15] stated that 128-bit input is a minimum requirement for block size. The information that needs to be secure for only minutes, hours, or perhaps weeks, a 64-bit symmetric key will suffice. For data that needs to be secure for years, or decades, a 128-bit key should be used. For data that needs to remain secure for the foreseeable future, one may want to go with as much as a 160-bit key [2]. Strength of Symmetric key encryption depends on the size of key used. For the same algorithm, encryption using longer key is harder to break than the one done using shorter key [16].

In this paper we tried to strength the security of Blowfish algorithm 64-bit by increase both block size and key length; Blowfish 128-bit increased the security of the original Blowfish 64-bit algorithm by increase key space up to 112 bytes instead of 56 bytes that leads to increase complexity of brute force attack as well as increase the block size to 128-bit.

The block cipher requires the generated cipher text to be distributed uniformly when dissimilar plaintext blocks are

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

used during encryption. By statistically analyzing the block cipher it can be determined if the tested algorithm meets this requirement. A non-random block cipher can be susceptible to attacks of many types [5].

The test suite [6] from NIST was selected for testing Blowfish 128-bit generated sequences. These statistical tests are consistent for estimating the generators of random and pseudo-random numbers that utilized in cryptographic applications. This attempt is considered an initial analysis of blowfish 128-bit, as no researcher has performed statistical tests on Blowfish 128-bit with ECB mode yet.

The five sections in this paper include the following: Section 2 describes the Blowfish 64-bit, Blowfish 128-bit and ECB mode; Section 3 categorizes and explains each Blowfish 128-bit Data type for statistical test; Section 4 provides the results of the experiment and empirical analysis of the randomness testing on Blowfish 128-bit with ECB mode, while Section 5 provides the conclusion and recommendations for future work.

2. THE BLOWFISH BLOCK CIPHER

Blowfish 64 is a symmetric block cipher that uses Feistel network, iterating simple encryption and decryption functions of 16 times. Each Feistel structure offers various advantages, particularly in hardware. In the decryption process of the cipher text, the only requirement is to reverse the key schedule. The BA can be divided into key expansion and data encryption ([4];[8-9] The key Expansion of BA begins with the P-array and S-boxes with the utilization of many sub-keys, which requires precompution before data encryption or decryption. The P-array comprises eighteen 32-bit sub-keys: P1, P2… P18.

In this section a maximum key of 448 bits are converted into several sub-key arrays of up to a total of 4168 bytes.

There are 256 entries for each of the four 32-bit S-boxes:

S1,0, S1,1,..., S1,255

S2,0, S2,1,..., S2,255

S3,0, S3,1,..., S3,255

S4,0, S4,1,..., S4,255

How these subkeys are calculated is explained below:

1. the P-array is initialized followed by the four S-boxes, with a fixed string, which has the hexadecimal digits of pi.

2. XOR P1 with the key's first 32 bits, XOR P2 with its second 32 bits, and so on, until the key's bits are up to P14. The cycle is iterated through the key bits until the entire P-array has been XOR-ed with key bits.

3. The Blowfish algorithm is then used to encrypt the all-zero string, employing the described subkeys in steps 1 and 2.

4. P1 and P2 are replaced with the step 3 output.

5. Encrypt the step 3 output with the Blowfish algorithm using the modified subkeys.

Replace P3 and P4 with the output of step 5.

The process is continued, and all elements of the P-array are replaced, followed by all four S- Boxes, with the output continuously changing.

Data encryption commences with a 64-bit block element of plaintext morphing into a 64-bit ciphertext. First the 64-bit segment is split into two equal segments that form the base of the BA. The next step is the implementation of the exclusive-or-operation (XOR) that is carried out between the first segment of the 32-bit block (L) and the first P-array. The 32-bit data obtained from step 2 is moved to the F function which permutes the data into a 32-bit block segment, which is XOR'ed with the second segment of the 32-bit block (R) of the 64-bit plaintext split. Upon completion of the XOR operation, the 32-bit segments, L and R are exchanged for future iterations of the BA. Figure1 illustrates the architecture of the Blowfish algorithm with 16 rounds. The input is an element of 64-bit data, X, which is divided into two 32-bit halves: XL and XR. Data Decryption is similar to Encryption data, but P1, P2... P18 are used in the reverse order.

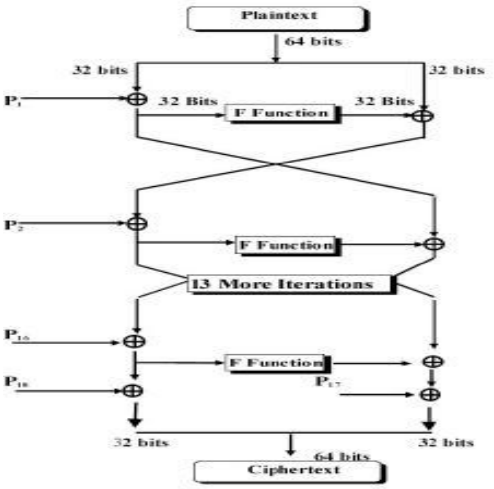


Figure 1. Blowfish Architecture

The F-Function of BA is probably the most complex part of the algorithm, as it is the only part that utilizes the S-boxes. It accepts a 32-bit stream of data and splits the data into four equal parts. Each 8-bit subdivision is changed into a 32-bit data stream using the corresponding of each subdivision S-box. The 32-bit data that is obtained is XOR'ed or combined to give a final 32-bit value for permutations of the BA (note that all the additions are modulo 2^{32}). Figure 2 Describes the architecture of the F function [4],10-12].

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

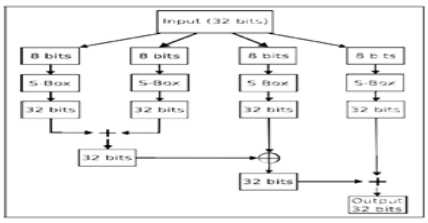


Figure 2. F-function Architecture

Blowfish 128-bit as the same outer structure of blowfish 64-bit is a Feistel network, iterating simple encryption and decryption functions of 16 times. Each operation in blowfish 128-bit was 64-bit instead 32-bit. In blowfish 128-bit the block size is 128 bits with variable key up to 112 bytes instead of 56 bytes. In expansion part the maximum key is 896 bits converted to the several subkey arrays up to a total of 8336 bytes instead of 4168 bytes.

There are 65536 entries for each of the four 64-bit S-boxes:
S1,0, S1,1,..., S1, 65535
S2,0, S2,1,..., S2, 65535
S3,0, S3,1,..., S3, 65535
S4,0, S4,1,..., S4, 65535

The Electronic Codebook (ECB) mode is a confidentiality mode. In this mode data is divided into blocks and each block is encrypted one at a time. Separate encryptions with different blocks are totally independent of each other. This means that if data is transmitted over a network or phone line, transmission errors will only affect the block containing the error. ECB is the weakest of the various modes because no additional security measures are implemented besides the basic algorithm. However, ECB is the fastest and easiest to implement [3].

The definition of the Electronic Codebook (ECB) mode is:
ECB Encryption:

$$C_j = CIPH_K(P_j)$$

ECB Decryption:

$$P_j = CIPH^{-1}_K(C_j), \quad \text{for } j = 1 \ldots n.$$

In ECB encryption, the forward cipher function is applied directly and independently to each block of the plaintext. The resulting sequence of output blocks is the ciphertext. While in ECB decryption, the inverse cipher function is applied directly and independently to each block of the ciphertext. The resulting sequence of output blocks is the plaintext. The ECB mode is illustrated in Figure 3[14].

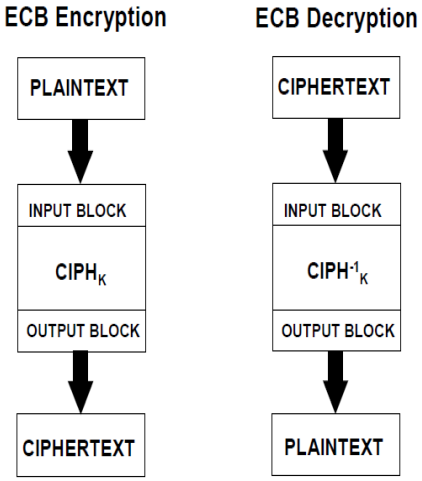


Figure3. the ECB mode

3.BLOWFISH 128 –BITS DATA TYPES

Testing the randomness on the Blowfish 128-bit was done by applying the NIST Statistical Suite [6]. All such testing consisting of 15 core statistical tests can be viewed as 188 statistical tests under different parameter inputs. Table 1 shows the Breakdown of the 188 statistics tests used in the experiments. In this section, we provide six Categories of Data such as, Random Plaintext/Random 128-Bit Keys, Low Density, High Density which are used to test the Advanced Encryption Standard (AES) candidate algorithms [7], Image, Text and Video files.

TABLE 1: BREAKDOWN OF THE 188 STATISTICAL TESTS APPLIED DURING EXPERIMENTATION

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Statistical Test	No. of P-values	Test ID
Frequency	1	1
Block Frequency	1	2
Cumulative Sum	2	3-4
Runs	1	5
Longest Run	1	6
Rank	1	7
FFT	1	8
Non Overlapping Template	148	9-156
Overlapping Template	1	157
Universal	1	158
Approximate Entropy	1	159
Random Excursions	8	160-167
Random Excursions Variant	18	168-185
Serial	2	186-187
Linear Complexity	1	188

3.1 Random Plaintext/Random 128-Bit Keys

The basis of this experiment is the data produced by the Blum-Blum-Shub (BBS) pseudo-random bit generator as it has been demonstrated to be a secure cryptographic pseudo-random bit generator and similar to the data type employed in testing Advanced Encryption Standard Finalist Candidates [7]. One hundred and twenty-eight sequences were constructed for the examination of the randomness of ciphertext (based on random plaintext and random 128-bit keys). Each sequence was due to the concatenation of 8128 ciphertext blocks of 128 bits (1040384 bits) using 8128 random plaintexts blocks of 128 bits and a random 128-bit key in ECB mode. BBS is implemented by using Java language (NetBeans IDE 7.2).

3.2 High-Density Plaintext

This experiment was base on data sets of 128 high density sequences. Each sequence consisted of 8257 ciphertext blocks, used different random 256-bit key per sequence. The first ciphertext block was calculated using an all ones plaintext block. Ciphertext blocks 2-129 were calculated using plaintext blocks consisting of a single zero and 127 ones, the zero appearing in each of the 128 bits positions of the plaintext block. Ciphertext blocks 130-8257 were calculated using plaintext blocks consisting of two zeros and 126 ones, the zeros appearing in each combination of two bit positions of the plaintext block [13].

3.3 Low-Density Plaintext

This experiment was base on data sets of 128 low density sequences. Each sequence consisted of 8257 ciphertext blocks. Used distinct a random 256-bit key per sequence. The first ciphertext block was calculated using an all zero plaintext block. Ciphertext blocks 2-129 were calculated using plaintext blocks consisting of a single one and 127 zeros, the one appearing in each of the 128 bits positions of the plaintext block. Ciphertext blocks 130-8257 were calculated using plaintext blocks consisting of two ones and 126 zeros, the ones appearing in each combination of two bit positions of the plaintext block [13].

3.4 Image Files

This experiment was based on data set of 128 sequences of image files in different formats. Each sequence was a result of the concatenation of 12290 (1573120 bits) ciphertext blocks of 128-bit using 12290 plaintexts blocks of 128-bit and a random 256-bit key in ECB mode.

3.5 Text Files

This experiment was based on a data set of 128 text files. Each file consisted of a sequence as a result of the concatenation of 8128 (1040384 bits) ciphertext block of 128-bit using 16256 plaintexts blocks of 128 bits and a random 256-bit key in ECB mode.

3.6 Video files

This experiment was based on a data set of 128 video files. Each file consisting of a sequence was a result of the concatenation of 8128 (1040384 bits) ciphertext block of 128-bit using 16256 plaintexts blocks of 64-bit and a random 256-bit key in ECB.

4. EXPERIMENTAL RESULTS

Testing the randomness on the Blowfish 128-bit was done on the six types of data mentioned in the previous section in both partially and full round considerations.

4.1. Full Round Testing (FRT)

When testing in full round with Blowfish128-bit, all six data types were generated, meaning that the data derived had to complete 16 for all types of data.

4.2 Partial Round Testing (PRT)

Soto and Bassham in [13] tested Twofish rounds in pairs. Twofish being a Feistel network, some of the data bits were left unaltered after each round and Twofish does not seem to be random under the test conditions. Nevertheless, after two rounds, there were effects on all data bits. Twofish rounds were measured in pairs, meaning, the rounds with even numbers from 2 to 14. Thus, in Partial Round Testing on Blowfish 128-bit, all six data types were generated using the Partial round of Blowfish128-bit in pairs from 2 to 14.

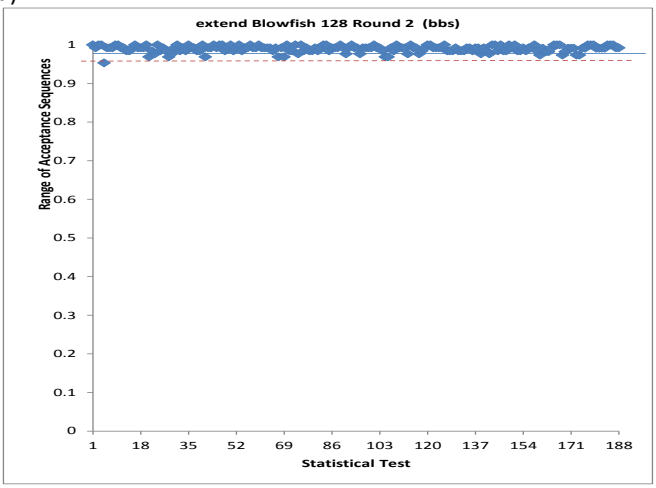
In the following we discuss the output of implementation Random Test for the six types of data on Blowfish 128-bit with ECB mode in PRT and FRT respectively.

1-Random Plaintext/Random 128-Bit Keys

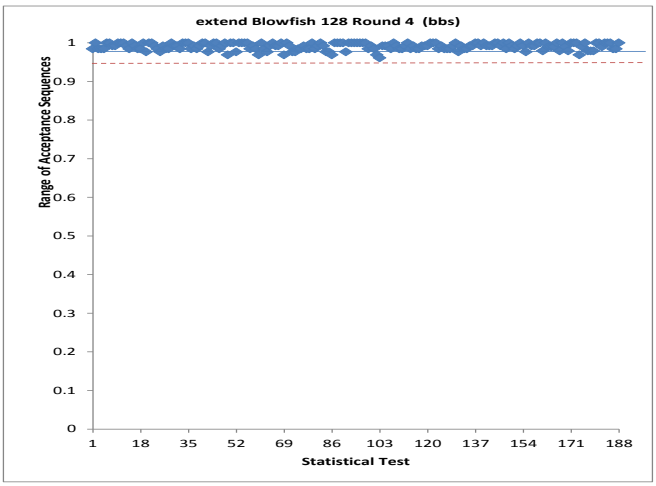
We illustrated the results of PRT and FRT of Blowfish 128-bit on this type of data with ECB mode in Fig 4. The dashed line in all Figures at 96.09% indicates the smallest proportion satisfying the 0.01 criterion of acceptance, while the solid line at 99% indicates the proportion expected.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

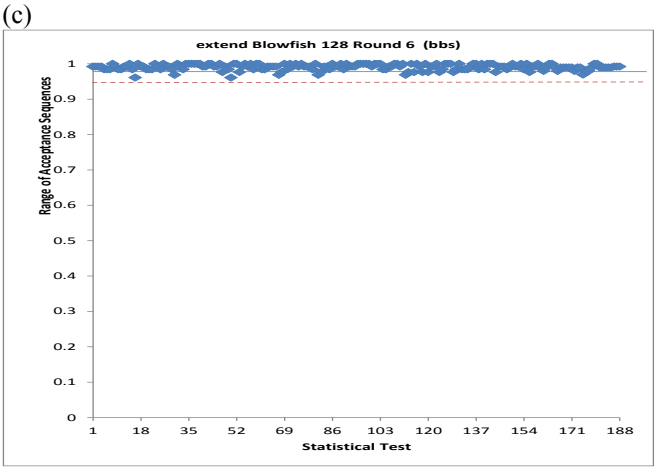
(a)



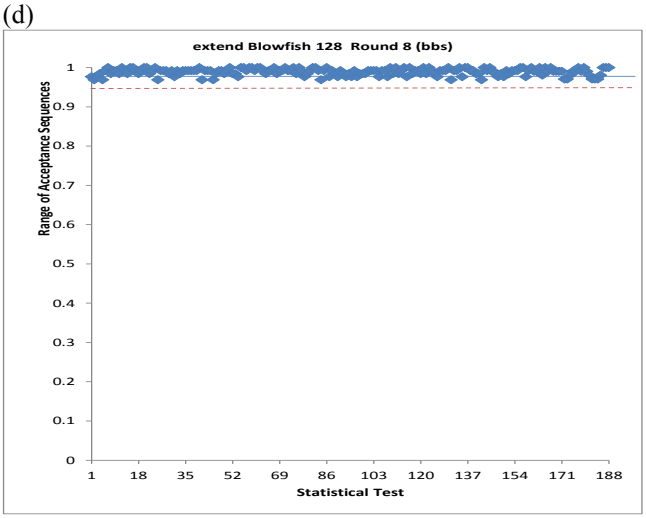
(b)



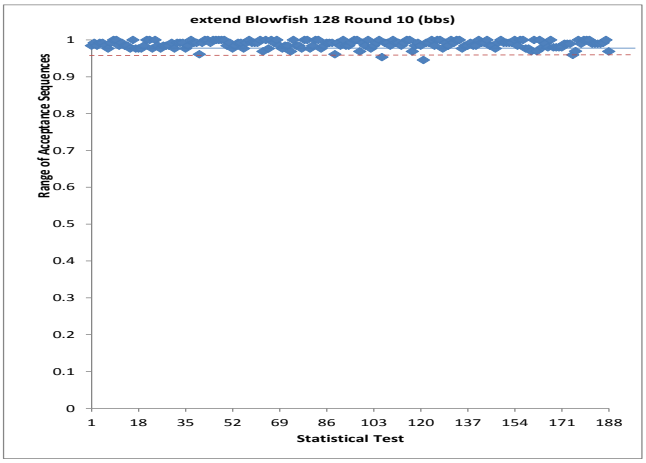
(c)



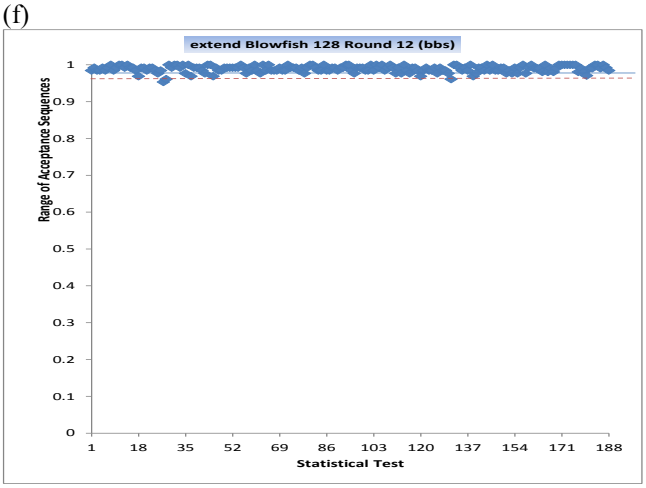
(d)



(e)

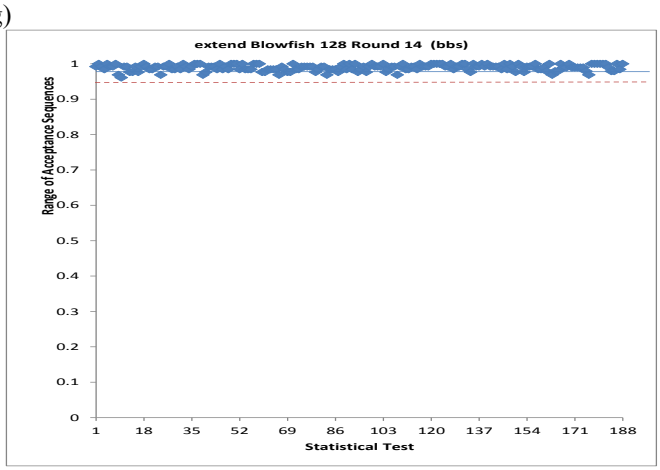


(f)



http://sites.google.com/site/ijcsis/
ISSN 1947-5500

(g)



(h)

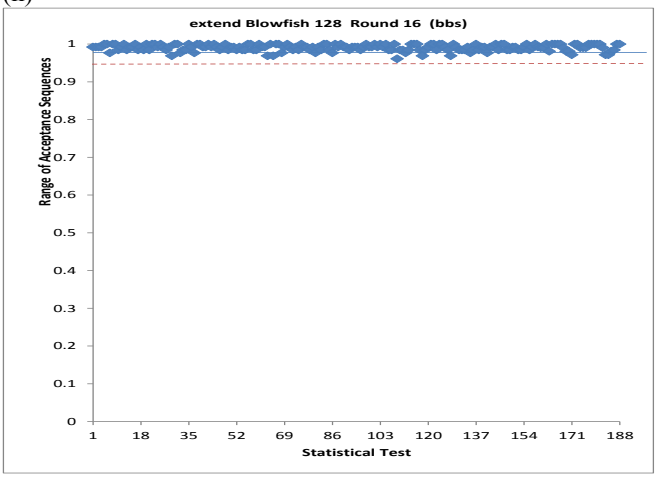
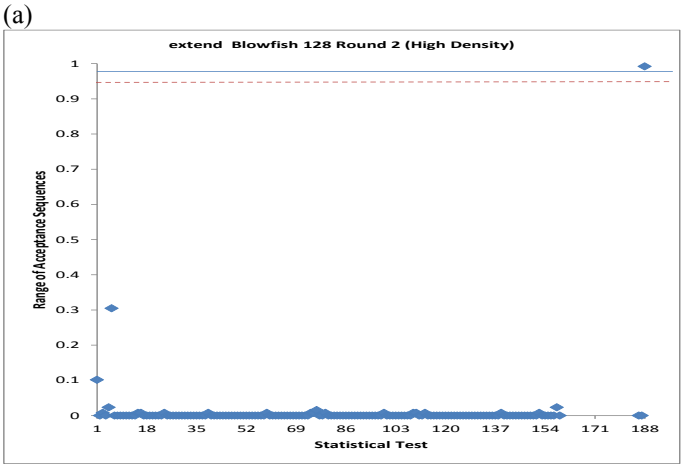


Figure 4 (a-h). Results of random plaintext/random 128-bit keys for (a) Second round with EBC mode, (b) Fourth round with EBC mode, (c) Sixth round with EBC mode, (d) Eighth round with EBC mode, (e)Tenth round with EBC mode, (f) Twelfth round with EBC mode, (g) Fourteenth round with EBC mode, (h) Sixteenth round with EBC mode.
It is evident that the output from the blowfish 128-bit on this type of data was random for all rounds because the most of the 188 statistical tests were above 96% algorithms.

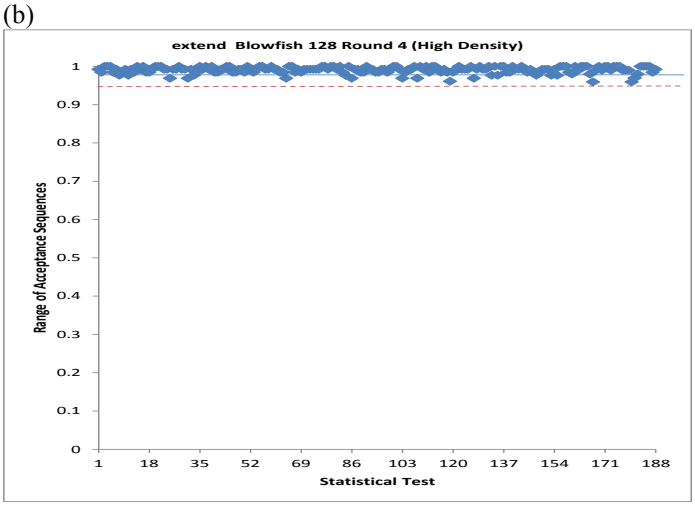
2-**High-density plaintext**

We illustrated the results of PRT and FRT of Blowfish 128–bit on this type of data with ECB mode in Figure 5 respectively.

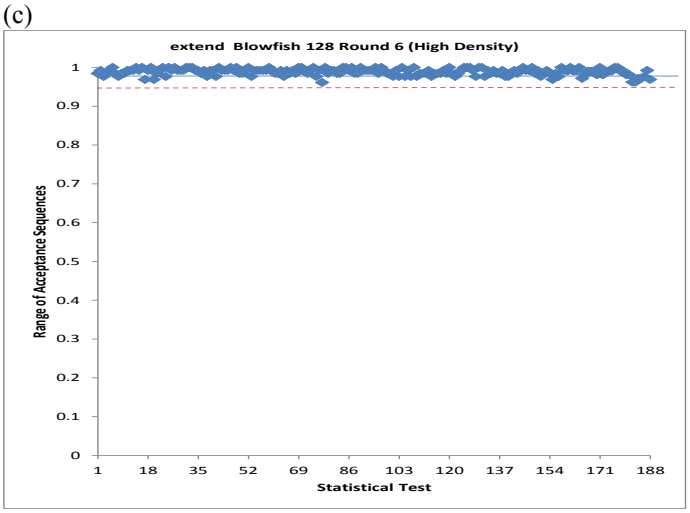
(a)



(b)

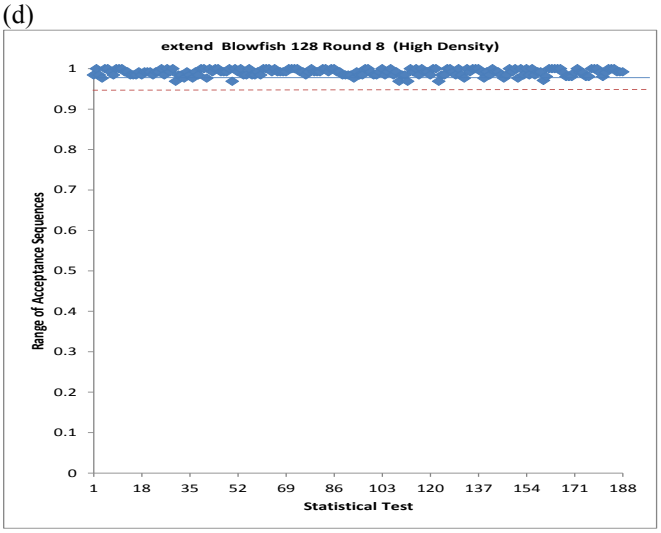


(c)

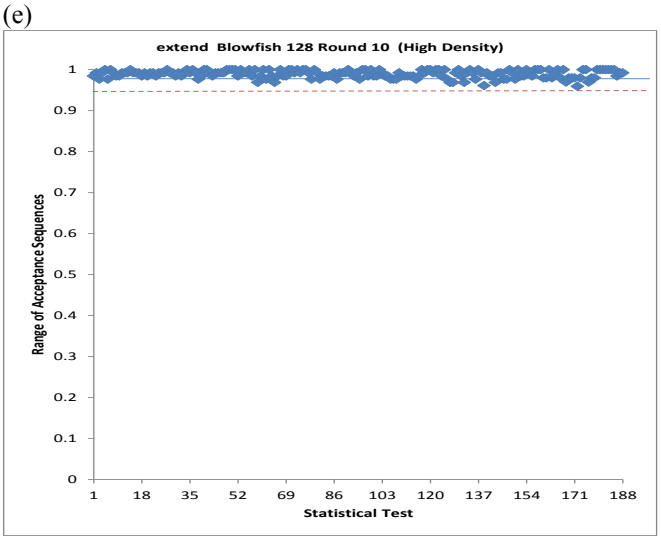


http://sites.google.com/site/ijcsis/
ISSN 1947-5500

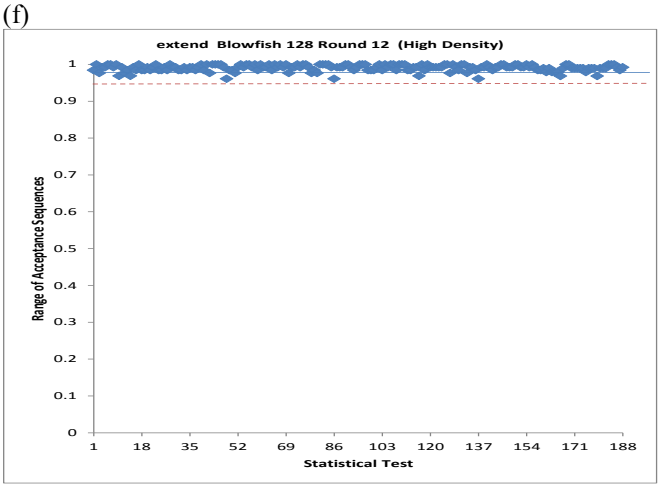
(d)



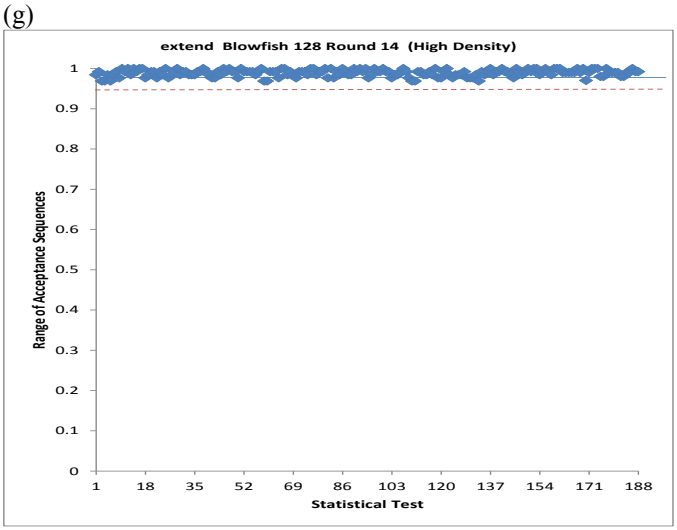
(e)



(f)



(g)



(h)

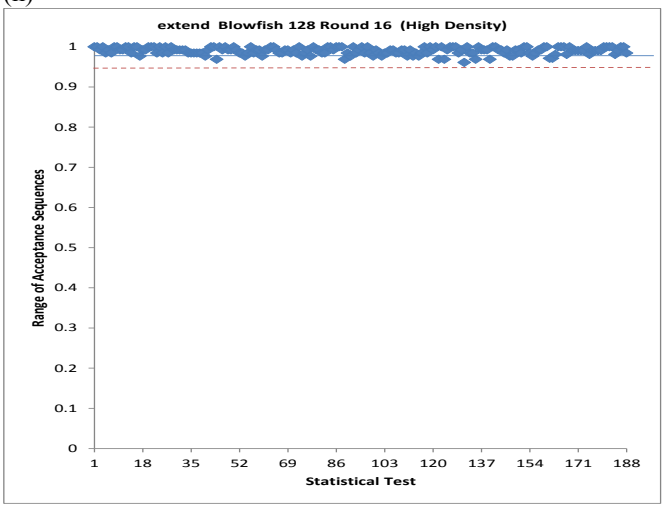


Figure 5 (a-h). Results of high density plaintext for (a) Second round with EBC mode, (b) Fourth round with EBC mode, (c) Sixth round with EBC mode, (d) Eighth round with EBC mode, (e)Tenth round with EBC mode, (f) Twelfth round with EBC mode, (g) Fourteenth round with EBC mode, (h) Sixteenth round with EBC mode.

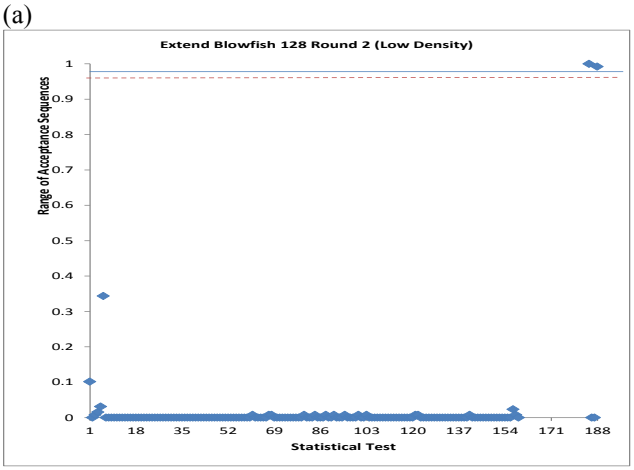
It is evident that the output from the Blowfish 128-bit still non-random by the end of the second round (the first round pair) because the majority of the 188 statistical tests were below 96%, it was random by the end of the fourth round (second pair round) because the majority of the 188 statistical tests were above 96%. Subsequent rounds produced similar statistics.

3-Low-Density Plaintext

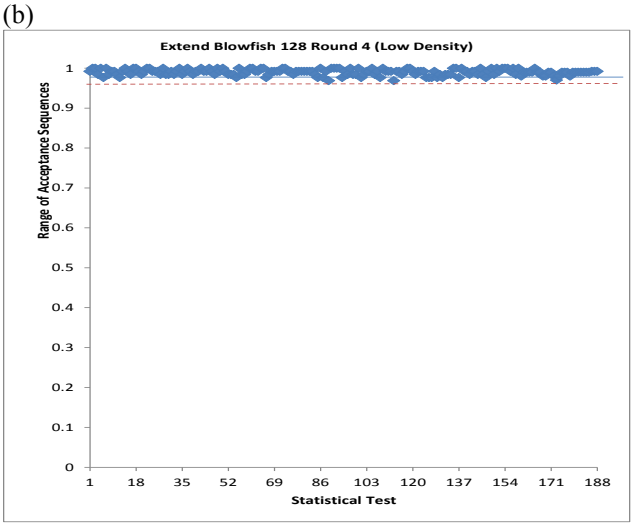
We illustrated the results of PRT and FRT of Blowfish 128–bit on this type of data with ECB mode in Figure 6 respectively.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

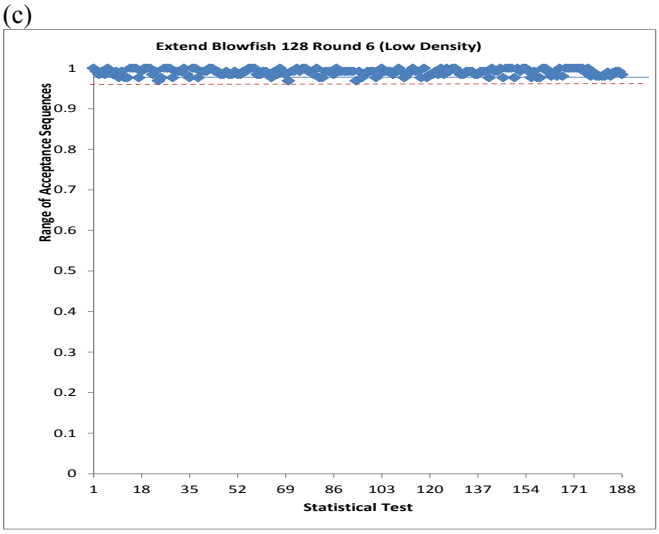
(a)



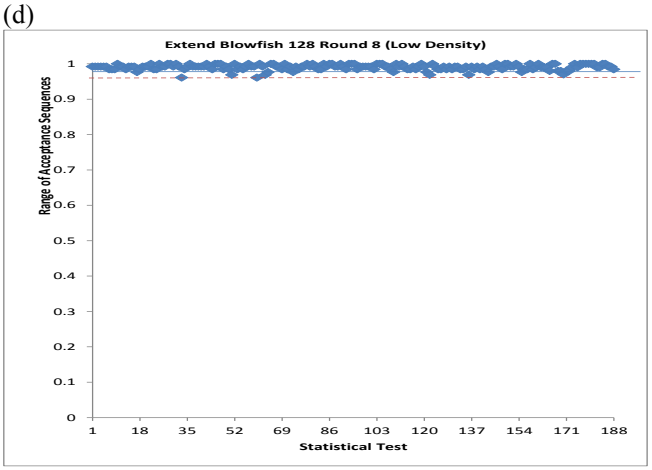
(b)



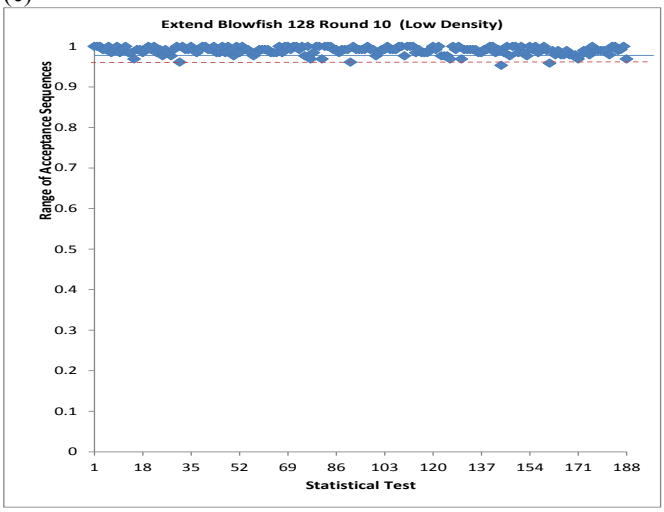
(c)



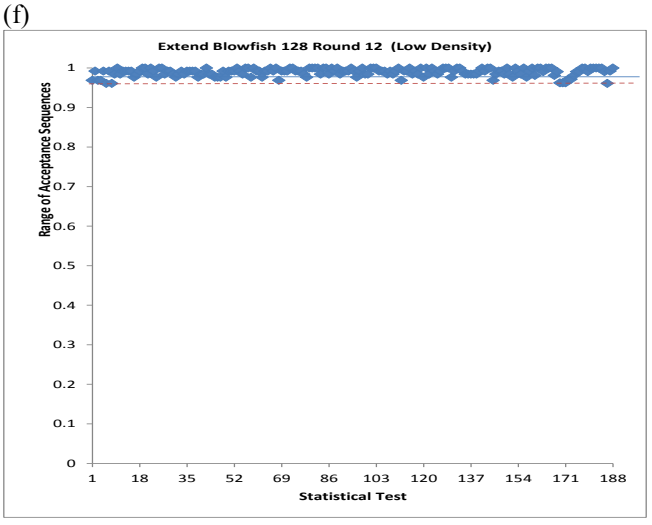
(d)



(e)

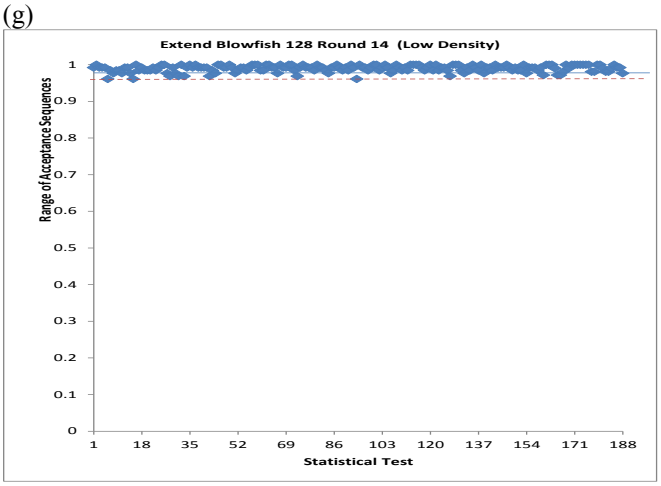


(f)



http://sites.google.com/site/ijcsis/
ISSN 1947-5500

(g)



(h)

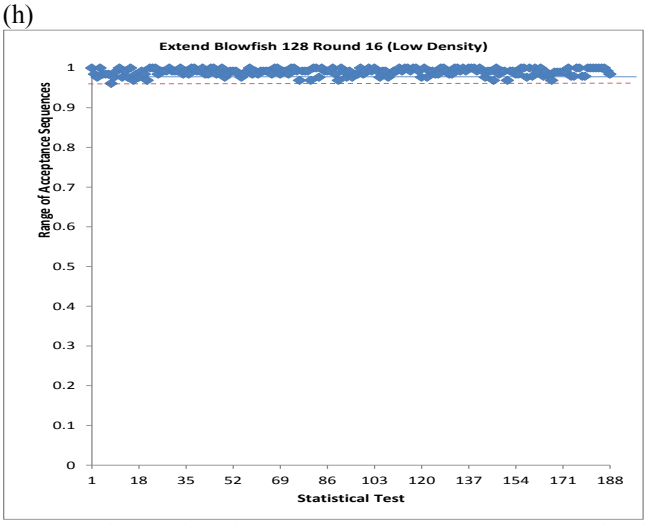


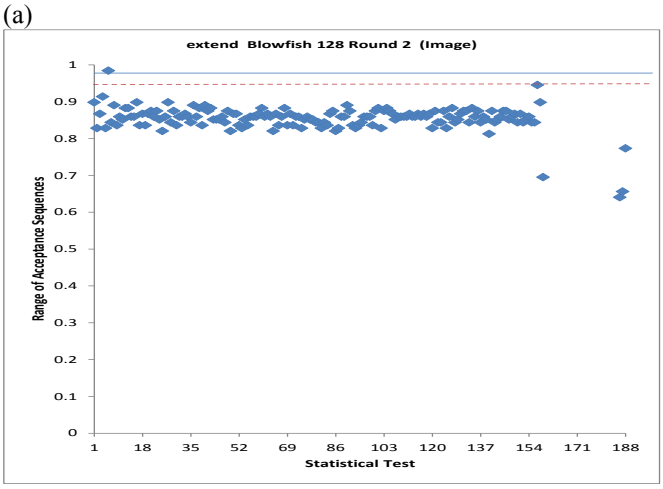
Figure 6 (a-h): Results of low density plaintext for (a) Second round with EBC mode, (B) Fourth round with EBC mode, (c) Sixth round with EBC mode, (d) Eighth round with EBC mode, (e)Tenth round with EBC mode, (f) Twelfth round with EBC mode, (g) Fourteenth round with EBC mode, (h) Sixteenth round with EBC mode.

It is evident that the output from blowfish 128-bit still non - random at the end of the second round (the first round pair) because the majority of the 188 statistical tests were below 96%, it was random by the end of the fourth round (second pair round) because the majority of the 188 statistical tests were above 96%. Subsequent rounds produced similar statistics.

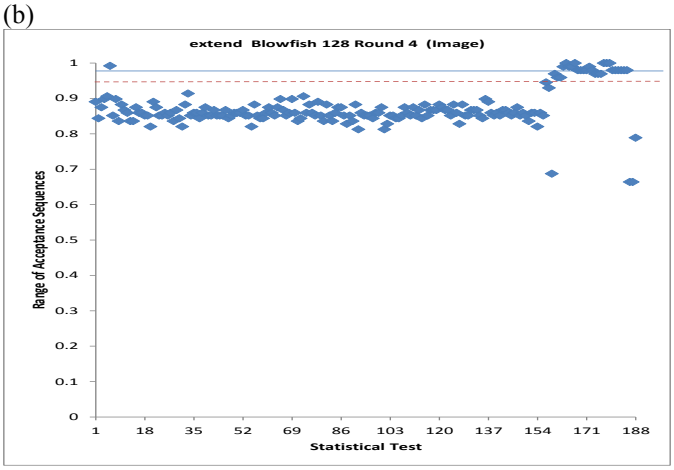
4-Image files

We illustrated the results of PRT and FRT of Blowfish 128–bit on this type of data with ECB mode in Figure 7 respectively.

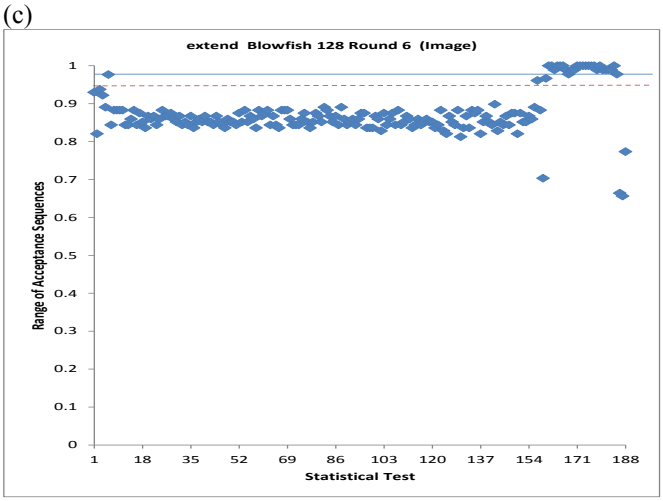
(a)



(b)

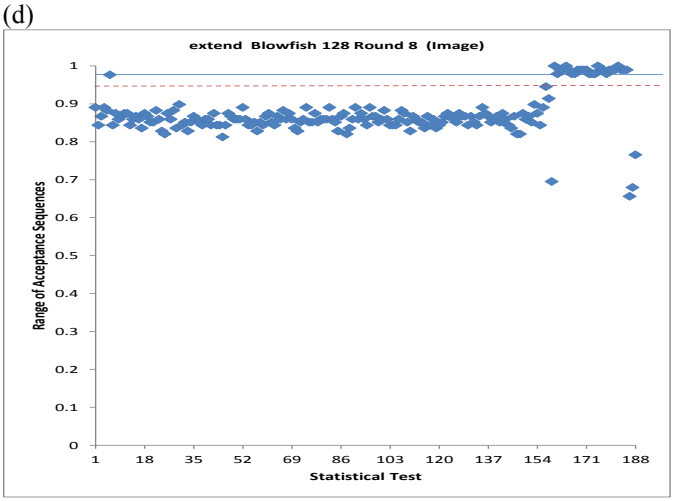


(c)

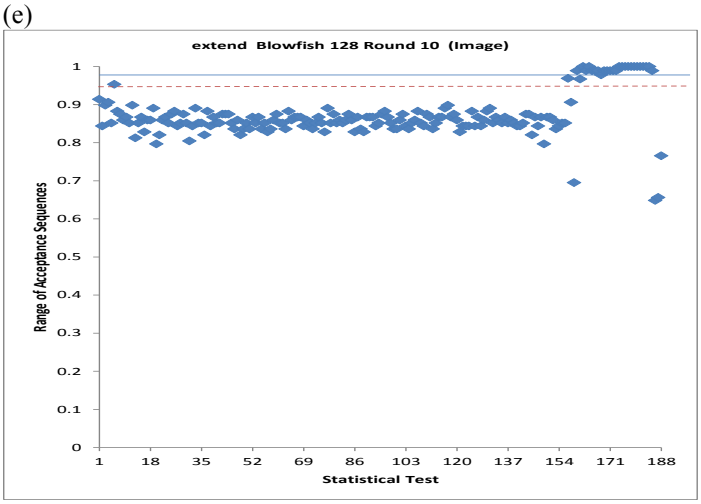


http://sites.google.com/site/ijcsis/
ISSN 1947-5500

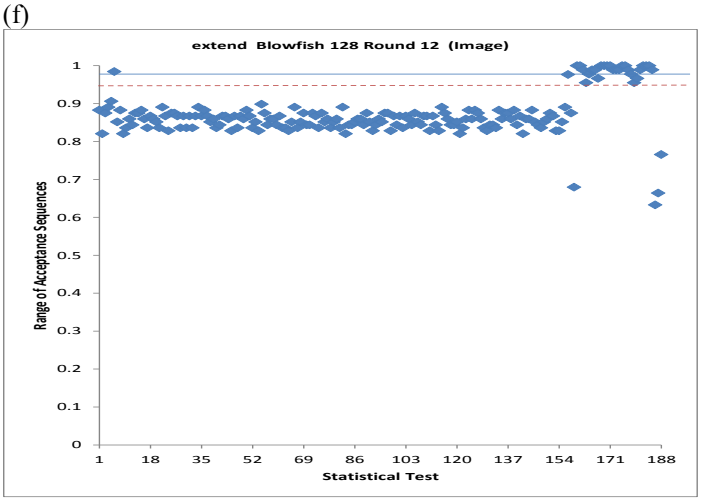
(d)



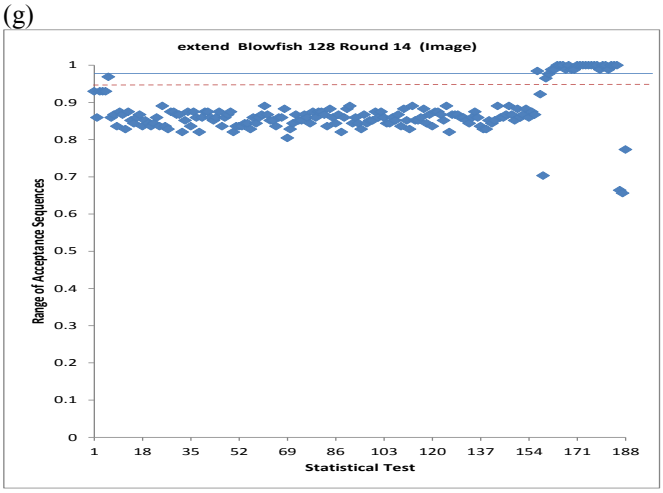
(e)



(f)



(g)



(h)

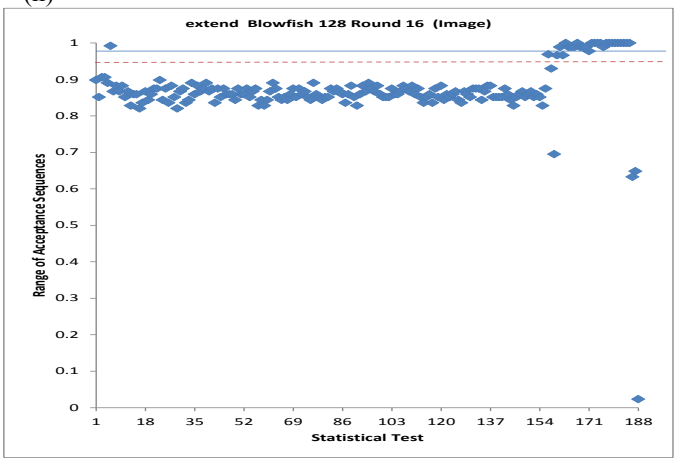


Figure 7 (a-h): Results of image files for (a) second round with EBC mode, (b) Fourth round with EBC mode, (c) Sixth round with EBC mode, (d) Eighth round with EBC mode, (e)Tenth round with EBC mode, (f) Twelfth round with EBC mode, (g) Fourteenth round with EBC mode, (h) Sixteenth round with EBC mode.

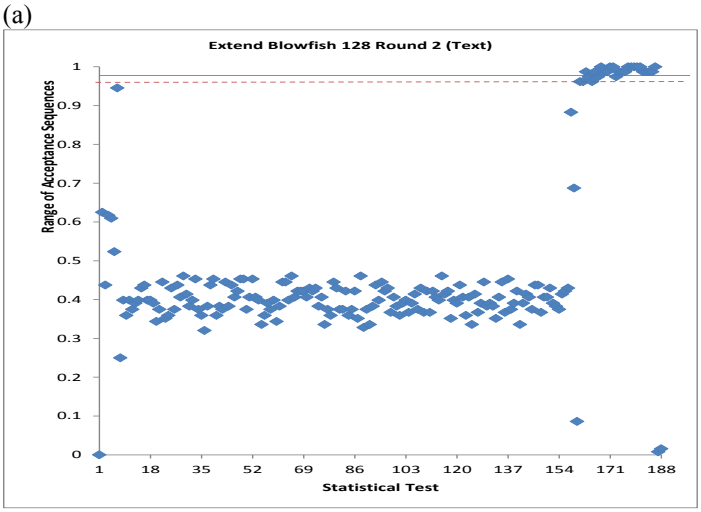
It is evident that the output from the blowfish 128-bit was non- random for all rounds because the most of the 188 statistical tests were below 96 %.

5-Text Files

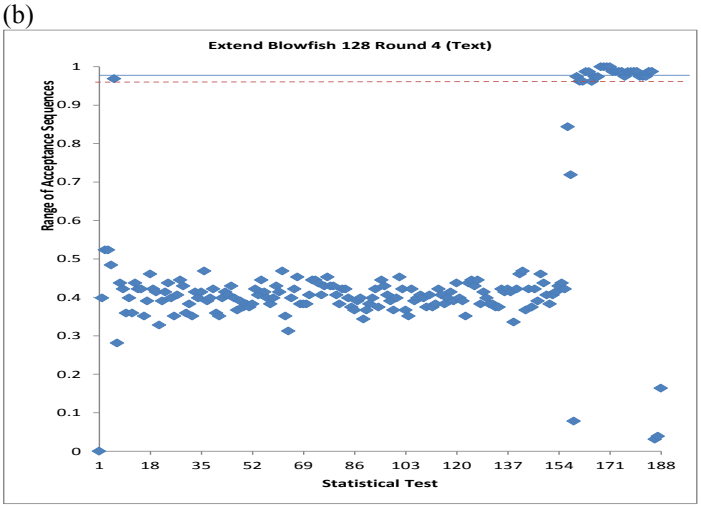
We illustrated the results of PRT and FRT of Blowfish 128–bit on this type of data with ECB mode in Figure 8 respectively.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

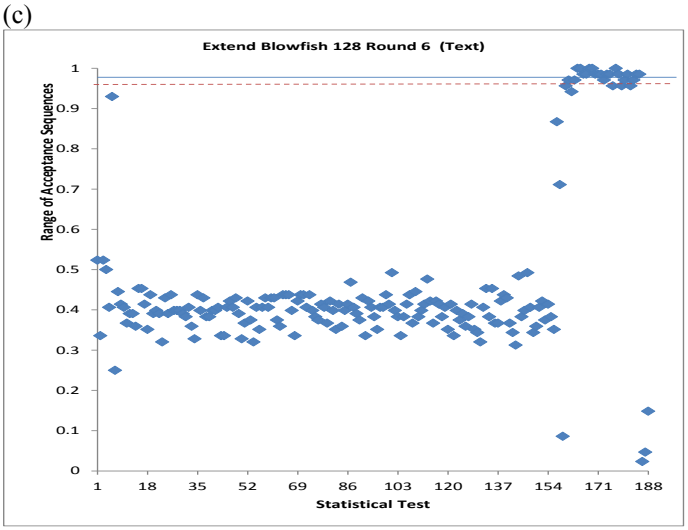
(a)



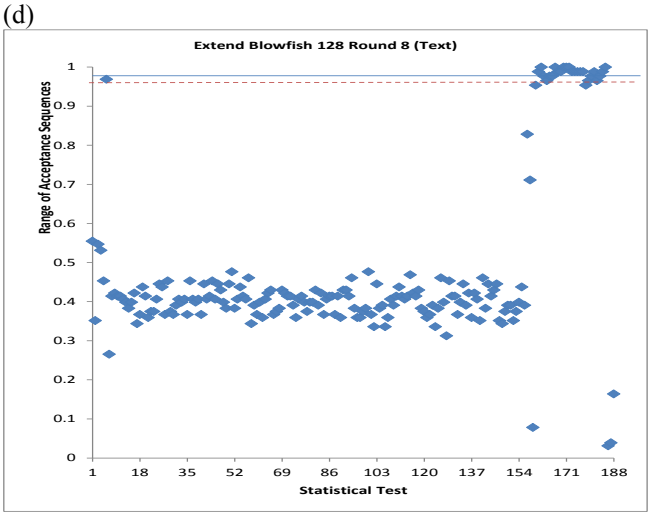
(b)



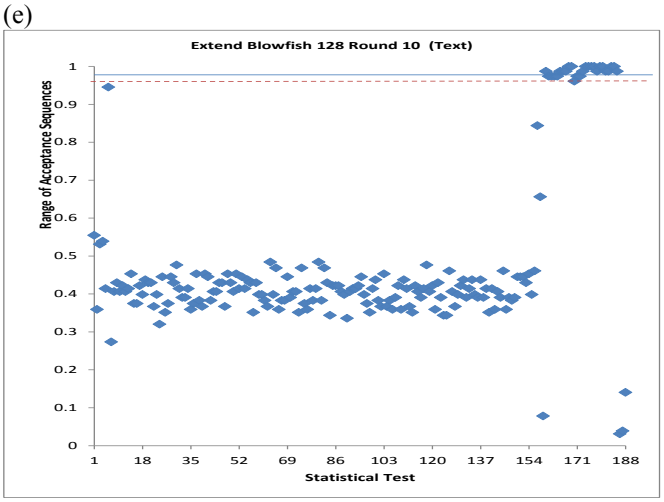
(c)



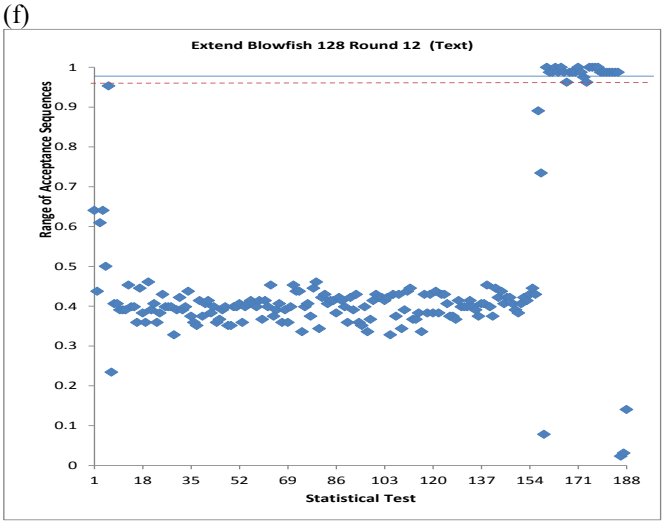
(d)



(e)

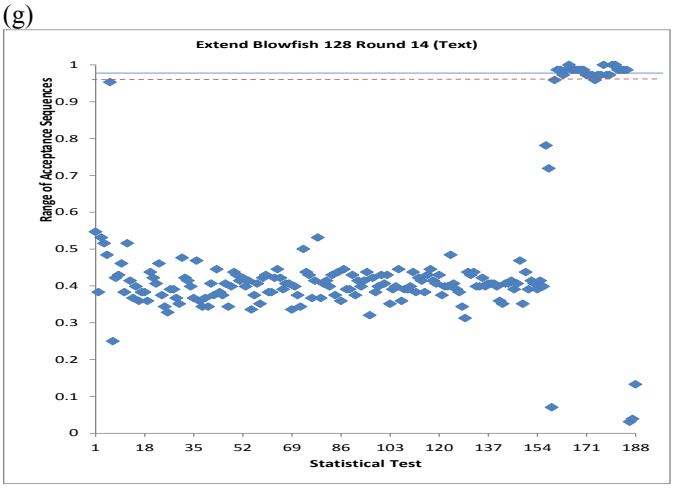


(f)



http://sites.google.com/site/ijcsis/
ISSN 1947-5500

(g)



(h)

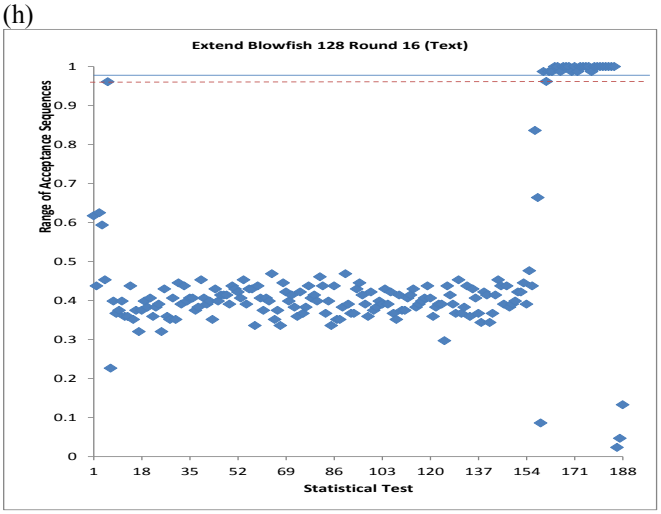


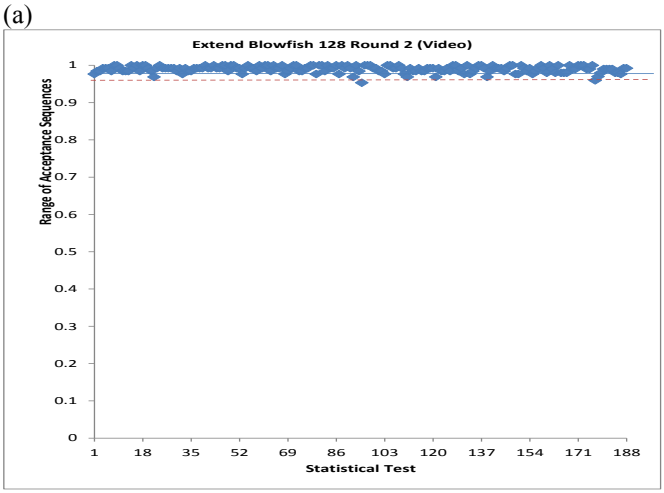
Figure 8 (a-h). Results of text files for (a) Second round with EBC mode, (b) Fourth round with EBC mode, (c) Sixth round with EBC mode, (d) Eighth round with EBC mode, (e)Tenth round with EBC mode, (f) Twelfth round with EBC mode, (g) Fourteenth round with EBC mode, (h) Sixteenth round with EBC mode.

It is evident that the output from the blowfish algorithm 128-bit was non- random for all rounds because the most of the 188 statistical tests were below 96 %.

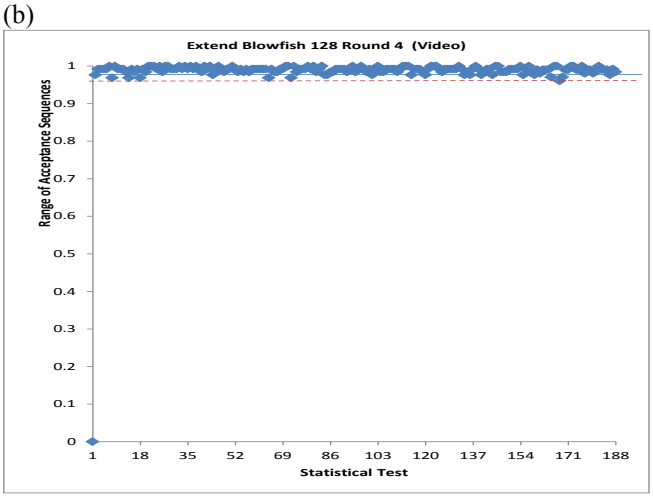
6-video Files

We illustrated the results of PRT and FRT of Blowfish 128–bit on this type of data with ECB mode in Figure 9 respectively.

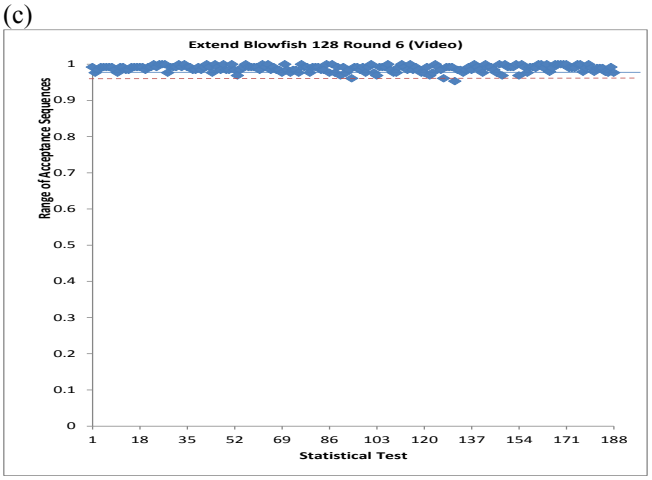
(a)



(b)

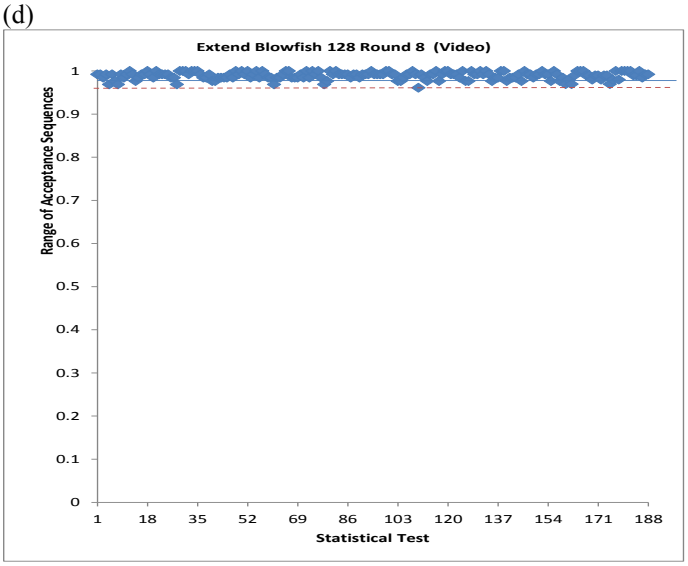


(c)

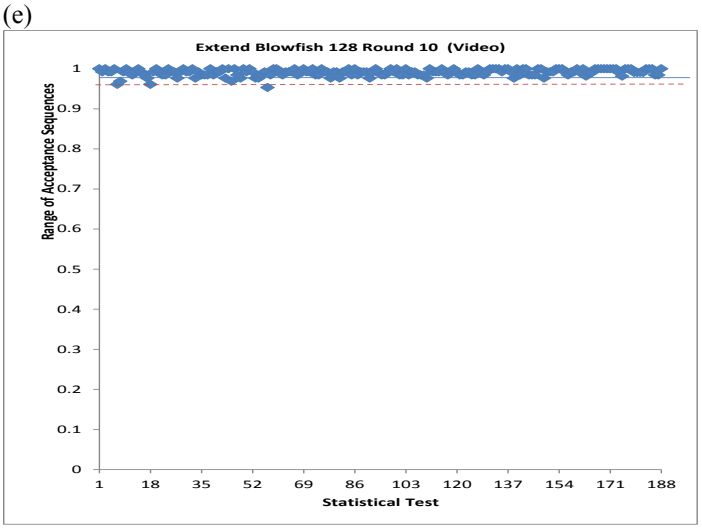


http://sites.google.com/site/ijcsis/
ISSN 1947-5500

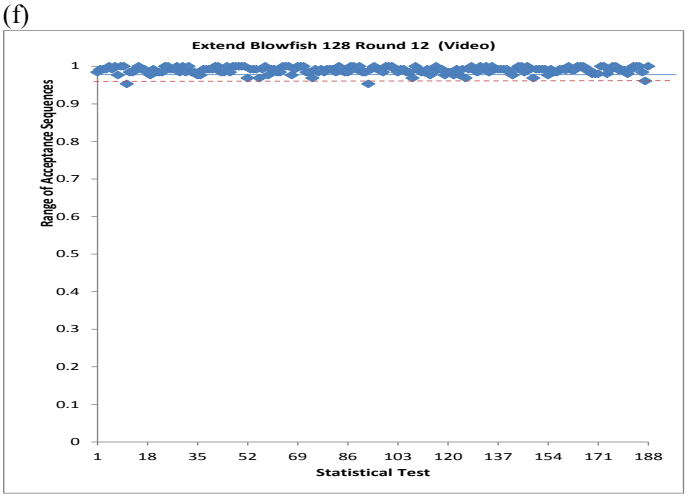
(d)



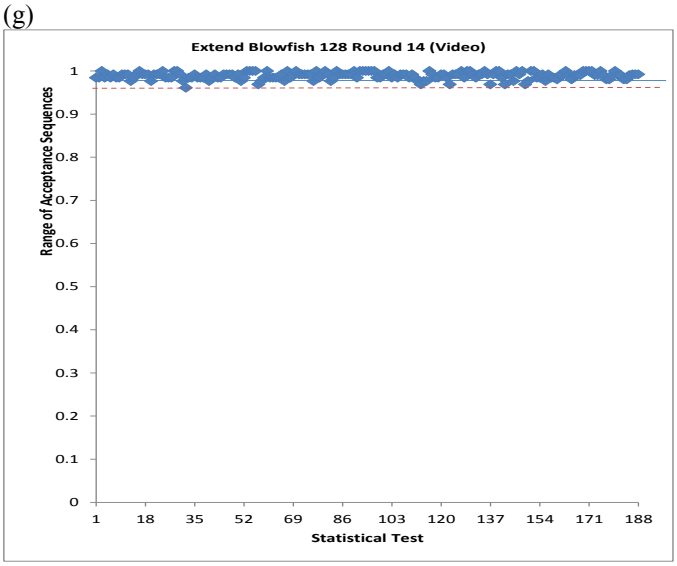
(e)



(f)



(g)



(h)

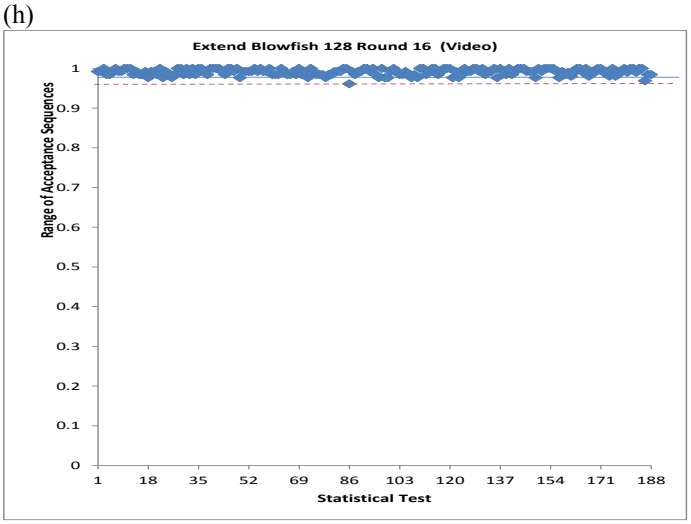


Figure 9 (a-h). Results of video files for (a) second round with EBC mode, (b) Fourth round with EBC mode, (c) Sixth round with EBC mode, (d) Eighth round with EBC mode, (e)Tenth round with EBC mode, (f) Twelfth round with EBC mode, (g) Fourteenth round with EBC mode, (h) Sixteenth round with EBC mode.

It is evident that the output from the blowfish 128-bit still non-random at the end of the first round because the majority of the 188 statistical tests were below 96%. It is random at the end of the second round because the most of the 188 statistical tests were below 96%. Subsequent rounds produced statistics similar.

5. Conclusion and Future Work

From the results, we can say the randomness of Blowfish algorithm 128-bit is better than blowfish 64-bit. Both

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

algorithms with ECB mode were not suitable with image and text files with large strings of identical bytes. But Blowfish 128-bit was better than Blowfish 64-bit with video files. Blowfish 128-bit is more secure against attacks than Blowfish 64-bit because key space and block size became double in Blowfish 128-bit that leads to increase complexity of brute attacks. Although this advantage it has distinct disadvantage that related with it need to the large memory. This finding in this paper can be considered the benchmark for starting point to investigate the effectiveness of the security of Blowfish 128-bit by enhancing its randomness as well as enhancing its performance by reduce memory requirement. Hence, in the future work we will focus on reduce memory requirements of blowfish 128-bit without compromising in security.

References

[1] MEYERS, R.K. AND A.H. DESOKY, *AN IMPLEMENTATION OF THE BLOWFISH CRYPTOSYSTEM*, IEEE INTERNATIONAL SYMPOSIUM ,SIGNAL PROCESSING AND INFORMATION TECHNOLOGY, ISSPIT, SARAJEVO, BOSNIA & HERZEGOVINA, 2008, PP 346-351.

[2] Moussa, A., *Data encryption performance based on Blowfish*, 47th International Symposium ELMAR, Zadar, Croatia,2005, pp 131-134.

[3]Thakur, J. and N. Kumar, DES, AES and Blowfish: Symmetric Key Cryptography Algorithms Simulation Based Performance Analysis, *International Journal of Emerging Technology and Advanced Engineering*, 2011, 1(2), pp 6-12.

[4] Cornwell, J.W. and G.A. Columbus, Blowfish Survey, 2012, pp1-6.

[5]Isa, H. and M.R. Z'Aba, *Randomness analysis on LED block ciphers*, Proceedings of the Fifth International Conference on Security of Information and Networks, Jaipur, India , 2012, pp 60-66.

[6] Rukhin, A., J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks , A. Heckert, J. Dray, S.Vo, *A statistical test suite for random and pseudorandom number generators for cryptographic application*s, National Institute of Standards and Technology Special Publication. Report number: 800-22, 2010, pp1-131.

[7] Soto, J. Randomness *Testing of the AES Candidate Algorithms*, National Institute of Standards and Technology. NIST IR 6390, September 1999, pp 1-pp9.

[8] Schneier, B., Description of a new variable-length key, 64-bit block cipher (Blowfish), in *Fast Software Encryption – Porceedings of the Cambridge Security Workshop*, Cambridge, United Kingdom, Lectures Notes in Computer Science 809, Springer-Verlag, 1994, pp. 191-204.

[9] Kumar, R.S., E.Pradeep, K.Naveen and R.Gunasekaran, A Novel Approach for Enciphering Data of Smaller Bytes, *International Journal of Computer Theory and Engineering*, 2(4), 2010, pp 654-659.

[10] Bagad, V.S. and I. Dhotre A. *Cryptography And Network Security*, ISBN 97788184313406 Second Revised Edition, Technical Publications Pune, India, 2008.

[11]Van Tilborg, H.C.A. and S. Jajodia(eds), *Encyclopaedia of cryptography and security*, ISBN 978-1- 4419-5906-5, 2nd edition, Springer Science and Business Media ,2011.

[12] Schneier, B., *Applied Cryptography: Protocols, Algorithms, and Source Code in C,* 2nd edition, New York: Wiley,1996.

[13] Soto, J. and L. Bassham, *Randomness Testing of the Advanced Encryption Standard Finalist Candidates*, National Institute of Standards and Technology. NIST IR 6483, 2000.

[14] Dworkin, M., *Recommendation for block cipher modes of operationmethods and techniques,National Institute of Standards and Technology Special Publicatio*n. Report number: 800-38A, 2001, pp 1-66.

[15] James Nechvatal, Elaine Barker, Lawrence Bassham, William Burr, Morris Dworkin, James Foti, Edward Roback(2000), Report on the Development of the Advanced Encryption Standard (AES), National Institute of Standards and Technology.

[16] Tingyuan Nie, Teng Zhang, 2009 A Study of DES and Blowfish Encryption Algorithm

[17] Alabaichi, A. M., Mahmood, R., Ahmad, F., & Mechee, M. (2013). Randomness Analysis on Blowfish Block Cipher Using ECB and CBC Modes.*Journal of Applied Sciences*, *13*, 768-789.

[18] ALabaich, A. M., Mahmood, R., & Ahmad, F. (2013). Randomness Analysis on Blowfish Block Cipher. *AWERProcedia Information Technology and Computer Science*, *4*(2).

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Relay Assisted Epidemic Routing Scheme for Vehicular Ad hoc Network

Murlidhar Prasad Singh
Deptt of CSE, UIT, RGPV Bhopal,
MP, India

Dr. Piyush Kumar Shukla
Deptt of CSE, UIT, RGPV
Bhopal,MP, India .

Anjna Jayant Deen
Deptt of CSE, UIT, RGPV Bhopal,
MP, India .

Abstract— **Vehicular ad hoc networks are networks in which no simultaneous end-to-end path exists. Typically, message delivery experiences long delays as a result of the disconnected nature of the network. In this form of network, our main goal is to deliver the messages to the destination with minimum delay. We propose relay assisted epidemic routing scheme in which we tend to use relay nodes (stationary nodes) at an intersection with a completely different number of mobile nodes which differs from existing routing protocols on how routing decision are made at road intersection where relay nodes are deployed. Vehicles keep moving and relay nodes are static. The purpose of deploy relay nodes is to increase the contact opportunities, reduce the delays and enhance the delivery rate. With various simulations it has been shown that relay nodes improves the message delivery probability rate and decreases the average delay.**

Keywords-VANETs, Routing Protocols, Relay Nodes.

I. INTRODUCTION

In Vehicular ad-hoc network (VANET), that is additionally known as SOTS (Self-organizing Traffic info System), may be a high-speed mobile outdoor communication network [1]. The essential idea of VANET is that vehicles within a particular communication range can exchange their information of speed, location and different knowledge obtained via GPS and sensors, and establish a mobile network automatically [2]. In this network every node acts as both a transceiver and a router, therefore multi-hop approaches are used to forward knowledge to further vehicle [3]. Compared with a traditional multihop, self-organizing networks without central nodes, there are many special features of VANET, together with e.g. short path life, the robust ability of computing and big storage, high-speed mobile nodes that lead to a fast modification of network topology, the flexibility of nodes to get power energy through vehicle engine. Additionally, nodes move in a very regular pattern, principally in single-way or two-way lane, with the features of one-dimensional, and therefore the vehicle track is mostly predictable [4]. VANET that is also called an opportunistic network designed to address several challenging connectivity issues such as sparse connectivity, long or variable delay, intermittent connectivity, asymmetric data rate, high latency, high error rates and even no end-to-end connectivity. The opportunistic network architecture adopts a store-and-forward paradigm and a common bundle layer located on the top of region-specific network protocols in order to provide interoperability of heterogeneous networks (regions). In this type of network, a source node originates a message (bundle) that is forwarded to an intermediate node (Relay or Mobile) thought to be closer to the destination node. The intermediate node stores the message and carries it while a contact is not available. Then the process is repeated, so the message will be relayed hop by hop until reaching its destination. The concept of opportunistic networking has been widely applied to scenarios like Vehicular ad hoc networks [5, 6].

In this paper, we exemplify the use of an opportunistic network to provide asynchronous communication between mobile nodes and relay nodes, on an old part of a city with a large area and restricted vehicular access. Mobile nodes (e.g., vehicles) physically carry the data, exchanging information with one another. They can move along the roads randomly (e.g. Cars). Relay nodes are stationary devices located at crossroads, with store-and-forward capabilities. They allow mobile nodes passing by to pickup and deposit data on them. We can also envision the possibility for the relay nodes to be able to exchange data with each other, and at least one of them may have a direct access to the Internet.

Some of the potential non-real time applications for this scenario are: notification of blocked roads, accident warnings, free parking spots, advertisements, and also gathering information collected by vehicles such as road pavement defects.

The rest of the paper has been organized as follows: Section II gives a brief description of related work. The proposed Relay Assisted Epidemic Routing is presented in section III. Performance comparison has been done in section IV. And finally section V includes concluding remarks and future works.

II. RELATED WORK

Multihop networking and relay technologies, where network nodes adopt self-organization properties and utilize short-range wireless communication to achieve network functionalities, have been actively researched in recent years. Routing is a critical issue in VANETs, and much work has been carried out. In flooding-based routing protocols, a data packet is to be disseminated to all nodes, including the packet's destination. A typical example is the classical epidemic routing (ER) [7]. The ER can achieve the least E2E packet delivery delay and optimal packet delivery ratio when

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the traffic load is light. However, it causes a lot of bandwidth waste and can lead to buffer overflow at intermediate vehicles when traffic load is heavy, thus degrading the routing performance.

Encounter based routing protocol MaxProp [8] forward data packets to nodes that have higher meeting probabilities with the packet destinations. This type of routing protocols takes advantage of the node movement pattern and uses such information to estimate the expected meeting time and probabilities between nodes. MaxProp[8] introducing a buffer management mechanism, which ranks the order of packets when exchanging among nodes and also being discarded when a buffer overflow occurs.

Auxiliary node assisted routing protocol use throwboxes or relay nodes to assist route selection and packet forwarding. In [9], Zhao et al. proposed to use throwbox to facilitate VANET routing. A throwbox is a device that can store and forward data. Furthermore, in [9], data exchanging is restricted to be carried out via throwboxes only, which often leads to low packet delivery ratio performance. The ParkCast protocol proposed in [10] suggests to use roadside parking cars as relays to help data disseminations in VANETs, whereas it does not discuss the details regarding how to use relays to help data forwarding. In [11], N. Banerjee et al. present an energy efficient hardware and software architecture for throwboxes. In [12], considers the cases where the throwboxes are fully disconnected or mesh connected, analyzing for each case the impact of the number of throwboxes over the performance of routing protocols. In [13], F. Farahmand et al. evaluates the relation of adding relay nodes to the overall network performance of a Vehicular Wireless Burst Switching Network (VWBS). It proposes and compares the performance of heuristic algorithms whose objective is to maximize the network performance in terms of delay or network cost. In [14], the author study the tradeoff of mobile networks enhanced by the deployment of relays, meshes and wired base station infrastructure. In [15-17], the authors study the impact of relay nodes on the performance of vehicular delay tolerant network and concluded that relay nodes gradually enhance the network performance by increasing the delivery rate and decreasing the average delay.

III. RELAY ASSISTED EPIDEMIC ROUTING PROTOCOL

A. Protocol Overview

Our proposed relay assisted epidemic routing (RAER) protocol contains a new mechanism rather than the mechanism used in the original epidemic routing protocol [7]. In epidemic routing when two nodes come within communication range they exchange messages and this process continues till message received successfully to the destination but in our proposed routing protocol we have deployed relay nodes at an intersection point by keeping in mind that intersection point may play a major role to find the best next hope.

Since the most effective path isn't continually available at the instant a packet reaches an intersection, we will deploy a relay node at every intersection to assist packet delivery. The relay node can store the packet for some time till the most effective path becomes available to deliver the packet. As illustrated in Figure.1 a packet is forwarded by wireless communication through vehicles A, B to the relay node R. Once the packet reaches R, the most effective path to deliver it's northward. However, there are no vehicles among communication range on this road at that time. Thus, R can store the packet for a short time, and forward it to the vehicle C. Once C passes the intersection and enters the northward road. From the figure, we can see that without the assistance of the static node, the packet is carried by B to the eastward road if B doesn't meet C at the intersection, which can result in a much longer packet's delivery path.

B. Protocol Description

Four kinds of information are needed to make the message forwarding decision:

(a) Set of vehicles currently in the communication range of Relay Nodes.
(b) The target destination of the message injected into a network.
(c) Gathering of expected traffic density information around relay node at a regular interval.
(d) Information regarding the message holding relay nodes or vehicles.

The algorithm of RAER protocols contains three stages:

(i) Vehicle to Relay Node:

Suppose a vehicle ñAò currently holding a message ñMò with target destination ñTò enters in an intersection, the vehicle needs to perform as follows:

(a) Ranking of all intersection points that are in communication range in increasing order on the basis of their end-to-end distance.
(b) Selection of best next hop intersection point on the basis of ranking information.

(ii) Relay Node to Vehicle:

In this case relay node need to make a decision on whether to forward message M to the current next hop vehicle or to wait till traffic density reaches to maximum limit.

(a) Forwarding a message to the first vehicle that are going towards the destination and delete it from relay node buffer.
(b) If no vehicle found that are going towards the destination till network density reaches to a maximum limit, deliver that message to any vehicle that are in communication range and farthest from it (relay node).

(iii) Vehicle to vehicle:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

In this case vehicle ñAò will forward message ñMò to any vehicle that are in communication range until it reaches its next intersection point or to the destination.

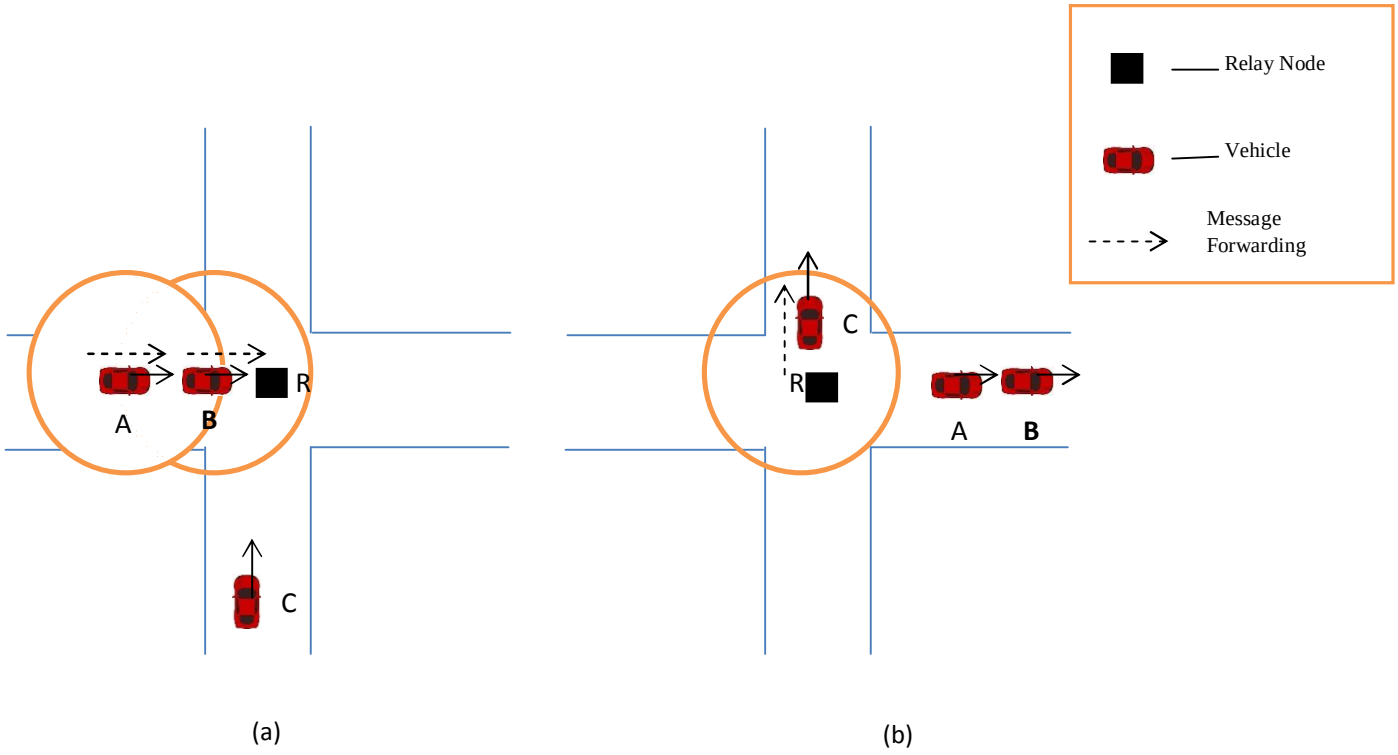


Figure 1: Relay Assisted Epidemic Routing in VANET

IV. PERFORMANCE EVALUATION

Performance Evaluation: We have used Opportunistic Network Environment (ONE) simulator [18] in order to evaluate the performance of Relay Assisted Epidemic Routing (RAER) protocol. One simulator is an open source simulator, designed by the Helsinki university and is freely available for research and development purpose.

A. Protocol Simulation Model

Here we have considered a scenario with mobile nodes i.e. vehicles. The detail of various simulation parameters is listed in the table 1.

B. Performance Metrics

The focus of the paper is to improve the message delivery performance considering two different types of cost metrics i. e. Both message delivery rate and message delay. These metrics are defined as follows:

1) Message Delivery Rate: The message delivery rate is defined as the ratio of the number of successfully delivered messages to total number of messages created at the source node. In mathematical form the delivery probability can be expressed as

Delivery probability =× Number of messages received / × Number of messages created

2) Average Message Latency/Delay: Average end to end delay can be defined as the average time taken from the source node to transfer the data packet to the destination node. It also includes all types of delay such as buffer delay, route discovery process, and delay during retransmission of the data packet, and propagation time etc. Only the data packets that successfully delivered to destinations that counted. In mathematical form the delivery probability can be expressed as

Message Delay =× (arrive time ï send time) / × Number of connections

The lower value of the end to end delay means better performance of the protocol.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Simulation Parameters	Simulation Parameter Values
Map Size	4500 m × 3400 m
Simulation Time	10000s
Buffer Size	30 MB
Packet Transmission Speed	250kBps
Number of Vehicles (with Fixed TTL=300 min)	20,40,60,80
Node Movement	Shortest Path Map Based Movement
Speed Vehicles	2.7-13.9m/s
Transmission Range	10m
Message Size	500kB-1MB
Message Generation Interval	25s-35s
TTL in Min(with fixed Vehicles=40)	20,40,60,80,100,150,200

Table 1. Simulation Environment Parameters

Here we have compared the Relay Assisted Epidemic Routing (RAER) with well known Flooding based routing protocol, one of which is MaxProp and the other is original Epidemic routing protocol. We have performed extensive simulation work with all these routing protocols in the same scenario setting with the above parameter and compared the performance of these routing protocols in terms of delivery probability rate and average latency/delay under different network sizes as well as different TTL.

A. Performance under different network size

We evaluate the impact of the node density by varying the number of mobile nodes. TTL is set to 300 minutes, and other parameters are same as Table 1. Figure.2 and figure 3 show the performance as the number of total mobile nodes varies from 20 to 80. Along with the increase of the number of mobile nodes, the delivery rates of all routings rise gradually. Overall, the RAER performs the best in terms of delivery rate and average delivery delay.

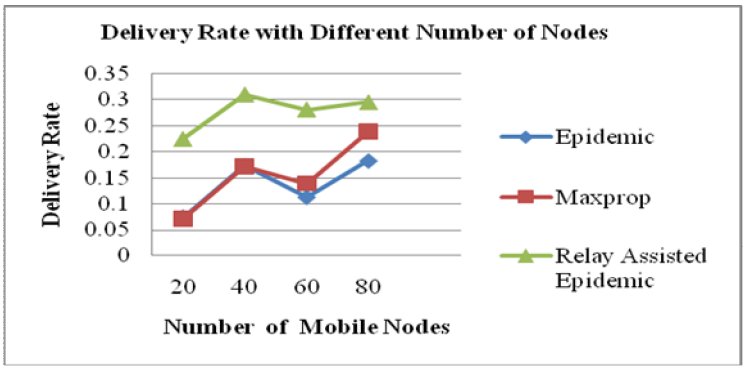


Figure 2 Delivery Rate with different Number of Nodes.

This is all due to the deployment of stationary relay nodes which enhance the contact opportunity of mobile nodes. As shown in Figure 2 and figure 3, the delivery rates rise, and an average delay of MaxProp and Epidemic first increases this is due to the sparse network where the number of mobile nodes is less hence less contact opportunities are there but in RAER routings delay constantly decreases as the number of total network nodes increases. The reason for this is that the opportunity for encountering between any two nodes increases as the network changes from sparse to dense, and this brings high delivery rates.

B. Performance under different TTL

Figure 4 and figure5 show the performance under different TTL. The figures show that the RAER performs better in comparison to the other routing protocols. As the TTL increases, the delivery rates of Epidemic and MaxProp routing rise gradually but in our scheme its slightly decreases initially and then remained unchanged and the average delivery delay increases slightly in all the routing scheme but RAER results minimum delay as compared to both.

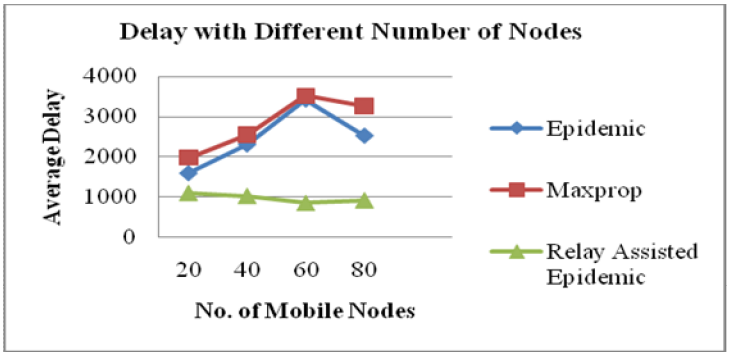


Figure-3 Delay with different Number of Nodes

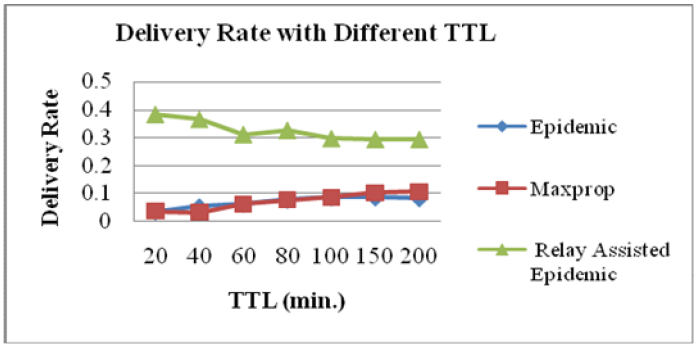


Figure 4 Delivery Rate with different TTL.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

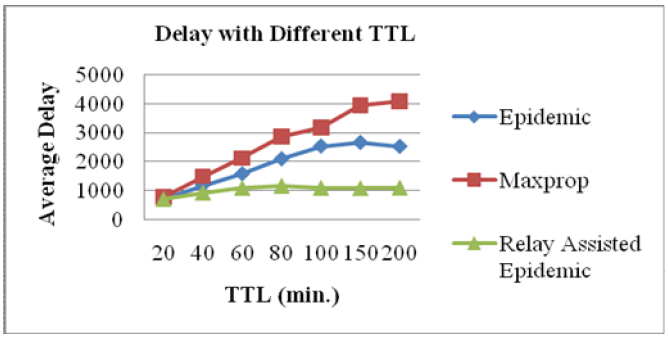


Figure 5 Delay with different TTL

V. CONCLUSION AND FUTURE WORK

As the vehicular network is completely composed of mobile nodes, the multi-hop data delivery performance may degrade under median or low vehicle densities where the network is frequently disconnected. In such cases, the performance may be improved by adding some stationary relay nodes at intersections to assist data delivery. Our proposed routing protocol RAER, a Relay Assisted Epidemic Routing protocol for Vehicular Ad hoc networks, makes the use of stationary relay nodes at intersection point which increases the contact opportunities among the mobile nodes leading to increase the delivery performance and it reduces the data delivery delay through when a packet reaches an intersection, it will be stored in the stationary relay node until the best delivery path becomes available to further deliver the packet. Our future work will lie on the way to designing relay node deployment strategy and evaluate the impact of relay node on proposed routing protocol.

REFERENCES

[1] Tom Van Leeuwen, Ingrid Moerman, Hendrik Rogier, Bart Dhoedt, Daniel De Zutter, Piet Demeester, ñBroadband Wireless Communication in Vehicles,ò *Journal of the Communications Network*, vol. 2, no.3, pp. 77-82, 2003. Article(CrossRef Link)

[2] ETSI TC ITS, TS 102 637-1, ñIntelligent Transport System (ITS); Vehicular Communication, Basic Set of Applications,ò *Part 1: Functional Requirements, Draft v2.0.3*, April, 2010.

[3] ETSI TC ITS, TS 102 636-1, ñIntelligent Transport System (ITS); Vehicular Communications, GeoNetworking,ò *Part 1: Requirements, v1.1.1*, March, 2010.

[4] ETSI TC ITS, TS 102 636-2, ñIntelligent Transport System (ITS); Vehicular Communications, GeoNetworking,ò *Part 2: Scenarios, v1.1.1*, March, 2010.

[5] W. Zhao, M. Ammar, and E. Zegura, "A Message Ferrying Approach for Data Delivery in Sparse Mobile Ad Hoc Networks," in *The Fifth ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2004)*, Roppongi Hills, Tokyo, Japan, May 24-26, 2004, pp. 187ï 198.

[6] A. Pentland, R. Fletcher, and A. Hasson, "DakNet: Rethinking Connectivity in Developing Nations," in *IEEE Computer*, vol. 37, 2004, pp. 78-83.

[7] A. Vahdat and D. Becker, "Epidemic Routing for Partially-Connected Ad Hoc Networks," Duke University, Technical Report CS-200006, Apr. 2000.

[8] Burgess J, Gallagher B, Jensen D, Levine B., ñMaxProp: routing for vehicle-based disruption-tolerant networks,ò In *Proceedings of IEEE INFOCOM'06*, Barcelona, Spain, April 2006; 1ï 11.

[9] Zhao W, Chen Y, Ammar M, Corner M, Levine B, Zegura E., ñCapacity enhancement using throwboxes in DTNs,ò In *Proceedings of IEEE MASS'06*, Vancouver,BC, Canada, October 2006; 31ï 40.

[10] Liu N, Liu M, Chen G, Cao J., ñThe sharing at roadside: vehicular content distribution using parked vehicles,ò In *Proceedings of IEEE INFOCOM'12*, Orlando, FL, March 2012; 2641ï 2645.

[11] N. Banerjee, M. D. Corner, and B. N. Levine, ñAn Energy-Efficient Architecture for DTN Throwboxes,ò in *INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, 2007, pp. 776-784.

[12] M. Ibrahim, A. A. Hanbali, and P. Nain, ñDelay and Resource Analysis in MANETs in Presence of Throwboxes,ò Performance Evaluation, vol. 64, pp. 933-947, October 2007.

[13] F. Farahmand, A. N. Patel, J. P. Jue, V. G. Soares, and J. J. Rodrigues, ñVehicular Wireless Burst Switching Network: Enhancing Rural Connectivity,ò In *The 3rd IEEE Workshop on Automotive Networking and Applications (Autonet 2008), Co-located with IEEE GLOBECOM 2008*, New Orleans, LA, USA, December 4, 2008.

[14] N. Banerjee, M. D. Corner, D. Towsley, and B. N. Levine, ñRelays, Base Stations, and Meshes: Enhancing Mobile Networks with Infrastructure,ò In *14th ACM International Conference on Mobile Computing and Networking (ACM MobiCom)*, San Francisco, California, USA, September, 2008, pp. 81-91.

[15] Vasco N.G.J. Soares, Farid Farahmand and Joel J.P.C. Radrigues, ñImproving Vehicular Delay-Tolerant Network Performance with Relay Nodes,ò In IEEE 2009.

[16] Mr. Mhamane Sanjeev C., Dr. Mr. Mukane S.M., ñ Impact of Relay Nodes on Performance of Vehicular Delay-Tolerant Network,ò In IJSRP Vol-2, Issues-7, July 2012 .

[17] Anand Mulasavavalgi and Anand Unnibhavi, ñ Impact of Relay Nodes on Vehicular Delay Tolerant Network,ò In IJCSC Vol-2, No.-2 July-2012.

[18] A. Keränen, J. Ott, and T. Kärkkäinen, ñThe ONE Simulator for DTN Protocol Evaluation,ò *In Proc. Intl. Conf. Simulation Tools and Techniques*, Mar. 2009.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Blind Optimization for data Warehouse during design

Rachid El mensouri and Omar El beqali
LIILAN/ GRM2I FSDM,
Sidi Mohammed Ben Abdellah University,
Fes, Morocco

ziyati Elhoussaine
RITM laboratory ENSEM - ESTC -
University Hassan II Ain chock
Casablanca, Morocco

***Abstract*—Design a suitable data warehouse is getting increasingly complex and requires more advance technique for different step. In this paper, we present a novel data driven approach for fragmentation based on the principal components analysis (PCA). Both techniques has been treated in many works [2][7]. The possibility of its use for horizontal and vertical fragmentation of data warehouses (DW), in order to reduce the time of query execution. We focus the correlation matrices, the impact of the eigenvalues evolution on the determination of suitable situations to achieve the PCA, and a study of criteria for extracting principal components. Then, we proceed to the projection of individuals on the first principal plane, and the 3D vector space generated by the first three principal components. We try to determine graphically homogeneous groups of individuals and therefore, a horizontal fragmentation schema for the studied data table.**

Keywords-component; data warehouse; optimization; PCA; vertical fragmentation; horizontal fragmentation; OLAP queries.

I. INTRODUCTION *(HEADING 1)*

Enterprise wide data warehouses are becoming increasingly adopted as the main source and underlying infrastructure for business intelligence (BI) solutions. Star schemes or their variants are usually used to model these applications. Queries running on such applications contain a large number of costly joins, selections and aggregations. To ensure a high performance of queries, advanced optimization techniques are mandatory. By analyzing the main optimization technique proposed in the literature we realize that some are applied when creating the schema of the data warehouse. We will focus on data partitioning [9], [13].

Especially the horizontal fragmentation problem is stated to be NP hard [2]. Roughly speaking, it is very difficult to find an optimal solution to problems in this class because of the fact that the solution space grows exponentially as the problem size increases. Although some good solutions for NP-hard [6] problems in general and the view selection problem in specific exist, such approaches encounter significant problems with performance when the problem size grows above a certain limit. More recent approaches use randomized algorithms in solving NP-hard problems.

II. RELATED WORK

Many research works dealing with horizontal partitioning problem were proposed in traditional databases and data warehouses. In the traditional database environment, researchers concentrate their work on proposing algorithms to partition a given table by analyzing a set of selection predicates used by queries defined on that table. Three types of algorithms are distinguished: minterm generation based approach [17], affnity-based approach [8] and cost-based approach [10]. Most of them concern only single table partitioning. Online analysis applications characterized by their complex queries motivate data warehouse community to propose methodology and effcient algorithms for horizontal partitioning. Noaman et al.[13] have proposed a methodology to partition a relational datawarehouse in a distributed environment, where the fact table is derived partition based on queries defined on all dimension tables. Munnekeetal.[12] proposed a fragmentation methodology for a multidimensional warehouse, where the global hypercube is divided into subcubes, where each one contains a sub set of data. This process is defined by the slice and dice operations(similar to selection and projection in relational databases). This methodology chooses manually relevant dimensions to partition the hypercube. In [7], a methodology and algorithms dealing with partitioning problem in relational warehouses are given. The methodology consists first in splitting dimension tables and using their fragmentation schemes to derive partition the fact table. It takes into account the characteristics of star join queries. Three algorithms selecting fragmentation schemes reducing query processing cost were proposed: genetic, simulated annealing and hill climbing. As in traditional database, these algorithms start with a set of predicates defined on all dimension tables. The main particularity of these algorithms is their control of generated fragments.[11] proposed an algorithm for fragmenting XML data warehouses based on k-means technique. Thanks to this k-means, the algorithms controls the number of fragments as in [7]. Three main steps characterize this approach: (i) extraction of selection predicates (simple predicates) from the query workload, (ii) predicate clustering with the k-means method and (iii) fragment construction with respect to predicate clusters. This approach is quite similar to affinity based approach [8]. To summarize, most the proposed works proposed in traditional databases are mainly concentrated on a

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

single table partitioning mode. Note that the most of studies in data warehouses are mainly concentrated on dependent table partitioning, where the fact table is partitioned based on the fragmentation schemes of dimension tables ,without addressing the problem of identification of dimension tables; except Munneke et al.'s work[12] that points out the idea of choosing and eliminating dimension to partition a hypercube.

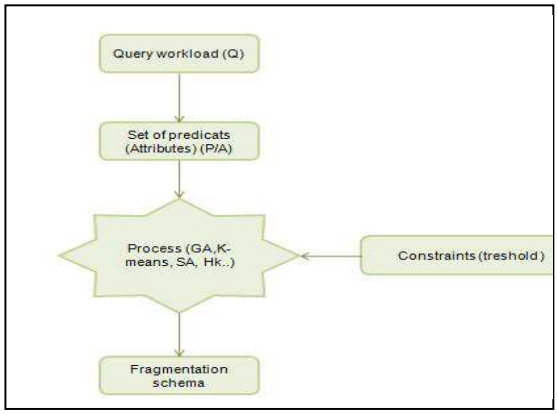


Figure 1. Requirement-Driven optimization schema

III. DATA DRIVEN APRROACH

The algorithm described bellow provides solutions without dependence with queries workload during design step, its focuses on data by considering data source.

Input: Data warehouse schema
1: Begin
2: PCA algorithm
3: define vertical fragment by component
4: define horizontal fragment by similarity (correlation) between tuples
8: End

However, the user may not know all the potential analysis required unlike requirement-driven approaches, it is able to propose new interesting multidimensional knowledge related to concepts already queried by the user. Finally, an optimized schema is derived from the input which support and mean fully ad-hoc queries

A. PCA Maintaining the Integrity of the Specifications

The Principal Components Analysis (PCA) is a part of the descriptive multidimensional methods called factorial methods [15]. We studied the classical PCA, for which all individuals have the same weight in the analysis and all variables are treated symmetrically. This can be problematic. The first criticism made by practitioners is that: if the old variables are heterogeneous, it is difficult to make sense to the principal components which are linear combinations of heterogeneous variables. The second criticism is that: if we change units on these variables, we can completely change the outcome of the PCA.

1) Decision based on the interpretation of the extracted components

Generally, the decision concerning the number of components to extract must also take into account the ability of researchers to interpret the extracted dimensions. There is no need to extract a component based on one of the last criteria if this component also defies any comprehension. Moreover, in 1996 and as in [7], Wood showed that an overestimation of the number of components was generally less damaging than an underestimation. The decision on the number of components to extract is difficult to make and has a significant share of subjectivity. It is suggested to compare the different criteria rather than applying directly the Kaiser criterion.

IV. USE CASE

To evaluate the optimization bought by the proposed algorithm as well as the impact of reducing time execution. Test is performed on a data warehouse which collects data from two databases in Oracle 10g. To achieve the normed PCA and to construct associated graphs with the software R-Revolution, we have used a machine with a CPU@2.80 Ghz and 4 GB of memory.

The star schema of our DW contains a fact table and three dimension tables.

TABLE I. THE BENCHMARK USED

Table	Attribute	
Journal (2 752 060 records)	Id_user	(fk)
	Id_nature_operation	(fk)
	Id_table	(fk)
	Num_operation	
	Agency_code	
	Data_source	
Table (176 records)	Id_table	(pk)
	Table_name	
User (103 records)	Id_user	(pk)
	User_name	
	User_entity	
Nature_operation (3 records)	Id_nature_operation	(pk)
	Nature_operation	

This study concerns only the fact table "Journal", the same approach can be applied to the other dimensions tables. In the first step, we are going to center and reduce variables to achieve a normed PCA. The study of correlations between the original variables and the principal components allow us to draw the circle of correlations and to determine graphically, under some conditions, candidate variables to be collected in vertical fragments.

By examining the correlation matrix, we notice that the presence of a strong correlation is between these three variables (num_operation, code_agence and data_source)

Identify applicable sponsor/s here. *(sponsors)*

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE II. CORRELATION MATRIX

	NAT_OP	AGENCY	USER	NAT_OP	TABLE	SOURCE
NUM_OP	1.000	-0.987	-0.012	0.030	0.038	0.993
AGENCY	-0.987	1.000	0.011	-0.023	-0.046	-0.993
USER	-0.012	0.011	1.000	0.026	-0.194	-0.013
NAT_OP	0.030	-0.023	0.026	1.000	0.089	0.033
TABLE	0.038	-0.046	-0.194	0.089	1.000	0.051
SOURCE	0.993	-0.993	-0.013	0.033	0.051	1.000

The determinant of this matrix is equal to 0.0001538589, its value is greater than 0.00001. So, it is a suitable situation to achieve the PCA according to Field (2000).

TABLE III. USEFUL DATA APPLIED TO PCA

Components	Eigenvalues	% of inertia	% of cumula-tive inertia
C1	2.988024690	49.80041	49,80041
C2	1.201345913	20.02243	69,82284
C3	1.018818662	16.98031	86,80315
C4	0.774942582	12.91571	99,71886
C5	0.012538419	0.2089737	99,9278337
C6	0.004329733	0.07216222	99,9999959
Total :	6.00000	100.00	

Table 2 shows the evolution of eigenvalues and inertia percentages of the fact table "Journal.", So we can reduce the data from 6 to 2 dimensions (C1,C2) keeping successfully 70% of the initial variance.

The third component C3 explains 1.02 units of variance, which corresponds to 16.98% of the total variance, and thus a total of 86.80% of variance explained by the first three components.

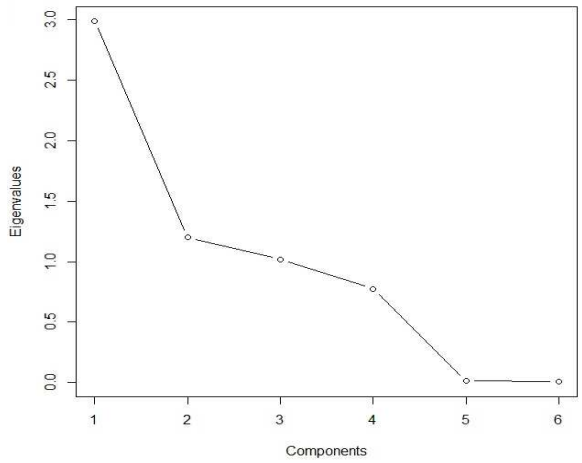


Figure 2. Evolution of eigenvalues.

1) Projection of individuals on the first principal plan

The first principal plan is spanned by the first two principal components; it keeps 70% of the total inertia contained in the data table.

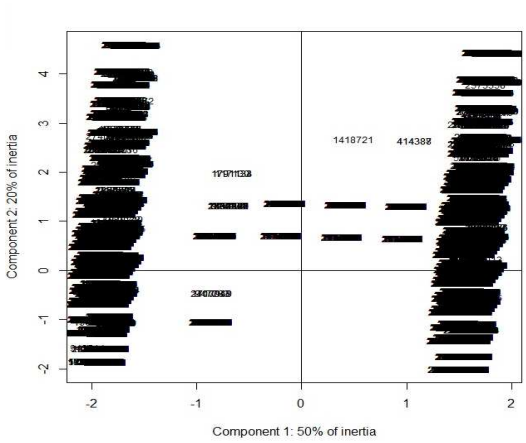


Figure 3. Projection of individuals on the first principal palne.

We note that individuals belong to two groups clearly distinguished: the first group consists of individuals with c1> = 0 and the second corresponds to those with c1 <0.

We can therefore propose a horizontal fragmentation schema that contains two fragments.

Then the projection of 2 752 060 individuals of our fact table on the 3D space returns to calculate the new coordinates of these vectors/individuals in the new basis of principal components. it is a heavy computation which was executed on a server host with 32GB of RAM in the Department of Mathematics & Computer in the Science Faculty of Oujda.

The vector space spanned by the first three principal components captures 86.80% of the total inertia contained in the data table.

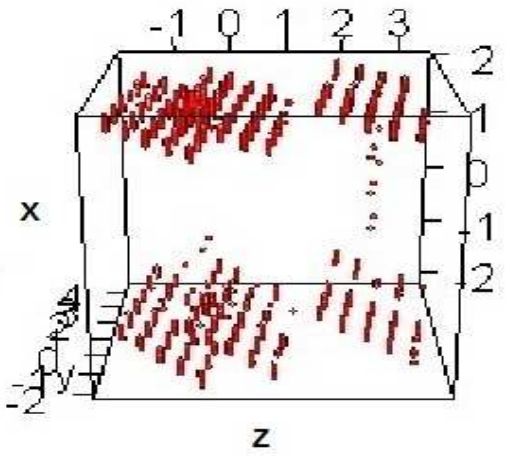


Figure 4. 3D projection of individuals.

In this projection, we are particularly interested to the position of individuals in relation to the third principal component (the Z axis), according to the previous Figure, it is

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

clear that this projection appears two blocks of individuals clearly separated relatively to the Z axis (Block1: individuals for which z <1.5 & Block2: individuals for which z >=1.5), that means additional horizontal fragments.

We can recommend a horizontal fragmentation schema that consists of the four following fragments:

- HF 1 : the set of individuals for which x>=0 & z >=1.5
- HF 2 : the set of individuals for which x>=0 & z < 1.5
- HF 3 : the set of individuals for which x< 0 & z >=1.5
- HF 4 : the set of individuals for which x< 0 & z < 1.5

In other hands by examining the correlation circle, we can say that the three variables (data_source, num_operation and agency_code) are well represented on the plan (C1, C2) because they are near to the edge of the circle.

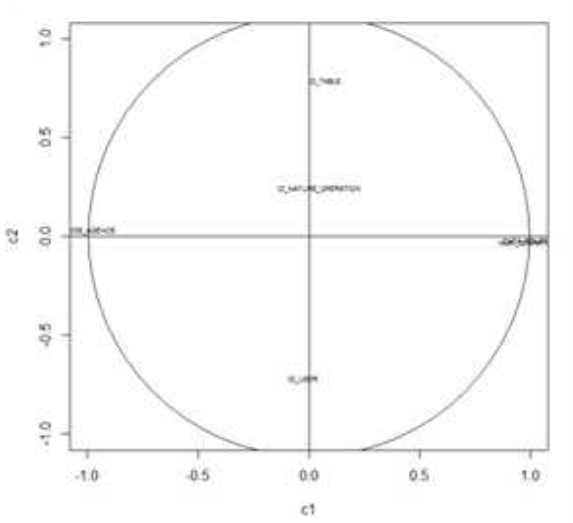


Figure 5. Correlation circle.

The two variables Data_source and Num_operation are highly correlated linearly and positively. While the variable Agency_code is highly correlated linearly and negatively with the two first variables.

We can therefore propose a vertical fragment for our DW composed by the three variables (Data_source, Num_operation and Agency_code).

The other three variables are not well represented (away from the edge of the circle), so we can not say anything about these variables. We notice that these variables have a correlation coefficient close to zero with the component C1.

Thus, we can recommend a vertical fragmentation schema that consists of the two following fragments:

- VF1: (data_source, num_operation, and agency_code).
- VF2: (id_table, id_user and id_nature_operation).

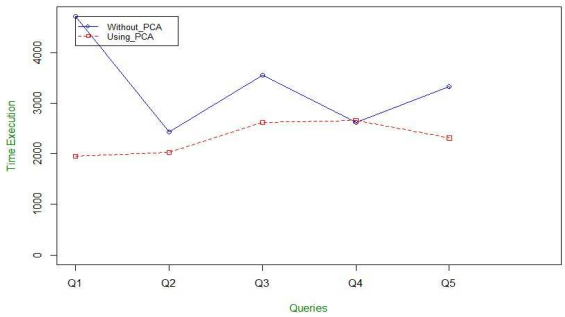


Figure 6. Queries execution time in different case.

However these results represent first tests about getting a new schema of fragmentation, more advanced analysis has to be done especially about OLAP queries workload. This allows to get a clearer idea about the impact of our idea figure 6 show the evolution, and the benefits of blind optimization in order to improve whole performance.

CONCLUSIONS

In this paper, we show the usefulness of using PCA in the context of fragmentation, and we have demonstrated this possibility for selecting horizontal and vertical fragmentation schema of data warehouses in order to reduce the execution time of OLAP queries.

The 3D projection vectors allowed us to highlight additional horizontal fragments invisible in 2D projection. However, a profound analysis of the representation quality concerning individuals/vectors and the interpretation of the new axes/components seems necessary to complete the study.

The next step of our work is to achieve the fragmentation on the Data Warehouse and to compare the execution time of a set of OLAP queries with and without mixed fragmentation based on our approach.

REFERENCES

[1] C. Duby, S. Robin, Analyse en composantes principales. *Institut National Agronomique Paris-Grignon. Département O.M.I.P*. 10 Juillet 2006.

[2] E. Ziyati, D. Aboutajdine and A. ElQadi, "Complete algorithm for fragmentation in data warehouse". *WSEAS in 7th WSEAS int. Conf. AIKED'08*. University of Cambridge. UK. Feb 20-22. 2008.

[3] H. F. Kaiser, "The application of electronic computers to factor analysis". *Educational and Psychological Measurement*, 20, 141-151, 1960.

[4] J. Horn, "A rationale and test for the number of factors in factor analysis". *Psychometrika, 30,* 179-185, 1965.

[5] J.M. Wood, D.J. Tataryn, and R.L. Gorsuch, "Effects of under and overextraction on principal axis factor analysis with varimax rotation". *Psychological Methods, 1*, 354-365. 1996.

[6] J.Yang, K. Karlapalem, and Q. Li. "Algorithms for materialize View Design in Data warehousing Environment", *Proceeding of 23rd VLDB Conference*, Athens, Greece 1997, P. 136-145.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[7] K. Boukhalfa, L. Bellatreche, and P. Richard, "Fragmentation primaire et dérivée: étude de complexité, algorithmes de sélection et validation sous Oracle10g". *LISI. Rapport de Recherche. N° 01 -2008*. Mars, 2008

[8] K. Karlapalem, S. B. Navathe, and M. Ammar. Optimal redesign policies to support dynamic processing of applications on a distributed database system. *Information Systems*, 21(4):353–367, 1996.

[9] L. Bellatreche, K. Boukhalfa, "La fragmentation dans les entrepôts de données : une approche basée sur les algorithmes génétiques", *Revue des nouvelles Technologies de l'information (EDA'2005)*, juin 2005, pages 141-160.

[10] L. Bellatreche, K. Karlapalem, and A. Simonet. Algorithms and support for horizontal class partitioning in object-oriented databases. *Distributed and Parallel Databases Journal*, 8(2):155–179, April 2000.

[11] H. Mahboubi and J. Darmont. Data mining-based fragmentation of xml data warehouses. *In ACM 11th International Workshop on Data Warehousing and OLAP (DOLAP'08)*, pages 9–16, 2008.

[12] D. Munneke, K. Wahlstrom, and Mohania M. K. Fragmentation of multidimensional databases. *in the 8th Australian Database Conference (ADC'99)*, pages 153–164, 1999.

[13] A. Y. Noaman and K. Barker. A horizontal fragmentation algorithm for the fact relation in a distributed data warehouse. *in the 8th International Conference on Information and Knowledge Management (CIKM'99)*, pages 154–161, November 1999.

[14] R. Bouchakri, L. Bellatreche, and K. Boukhalfa, "Administration et tuning des entrepôts de données : optimisation par index de jointure binaires et fragmentation horizontale," *Doctoriales STIC'09*. Msila. Décembre 2009.

[15] R. Elmansouri, E. Ziyati, D. Aboutajdine, and O. Elbeqqali, "The fragmentation of datawarehouses: an approach based on principal components analysis". *ICMCS'12. Tanger*. 10-12 Maï 2012.

[16] R. B. Cattell, "The scree test for the number of factors". *Multivariate Behavioral Research*, 1, 245-276, 1966.

[17] M. T. ¨ Ozsu and P. Valduriez. Principles of Distributed Database Systems : *Second Edition. Prentice Hall*, 1999.

AUTHORS PROFILE

Rachid elmensouri is currently an engineer at Alomrane organism in Oujda, preparing his PHD in Computer Science at University Sidi Med Ben AbdEllah, Fès, Morocco.

Omar El Beqqali is currently Professor at Sidi Med Ben AbdEllah University. He is holding a Master in Computer Sciences and a PhD respectively from INSA-Lyon and Claude Bernard University in France. He is leading the 'GRMS2I' research group since 2005 (Information Systems engineering and modeling) of USMBA and the Research-Training PhD Unit 'SM3I'.

Ziyati Elhoussaine is a Professor, in Computer Engineering department in ESTC Institute of Technology, Casablanca, Morocco, received PHD degree in Computer Science from Mohammed V. University in 2010, presently, his domain area is Business Intelligence, Big Data and Data warehousing.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Traffic Intensity Estimation on Mobile Communication Using Adaptive Extended Kalman Filter

Rajesh Kumar
Department of Electronics and Communication
Jabalpur Engineering College
Jabalpur, India (482011) .

Agya Mishra
Department of Electronics and Communication
Jabalpur Engineering College
Jabalpur, India (482011) .

Abstract— **Traffic estimation is an important task in network management and it makes significant sense for network planning and optimization. The biggest challenge is to control congestion in the network and the blocking probability of call to provide users a barrier less communication. Effective capacity planning is necessary in controlling congestion and call drop rates in mobile communication thus an accurate prediction of traffic results congestion control, effective utilization of resources, network management etc. In this paper a real time mobile traffic data of different mobile service providers are estimated using adaptive Extended Kalman Filter method. This paper evaluates compares and concludes that, this approach is quite efficient with min normalized root mean square error (NRMSE) and it can be used for real time mobile traffic intensity estimation.**

Keywords—Traffic Intensity Estimation, recursive filter, Mobile Traffic data, Extended Kalman filter, NRMSE.

I. Introduction

Traffic intensity estimation play increasingly important role for optimization on mobile networks and capacity planning. In today's cellular telecommunications systems, network resources are allocated according to the peak traffic load. So an accurate prediction plays an important role in resource allocation, capacity planning etc. A wireless cell can host limited number of calls, so in this condition when a new user will arrive and request cell to connect for the new call or request for the hand off so there will be a chance that user will either not be able to connect for the new call or the call drop may takes place in case of hand off request. So to avoid this accurate prediction is required to provide effective utilization of resources, network management, capacity planning etc. Capacity planning may involve monitoring the network traffic congestion and when the blocking probability increases then sufficient resources are allocated to cope up with the condition of network traffic congestion .This paper presents estimation of mobile traffic based on adaptive Extended Kalman filer method**.** Estimating mobile traffic in erlang is basically estimating the offered load. It is given in reference [2] that the blocking probability of a call is directly related with the offered load i.e. if the offered load increases than there will be a chance that blocking probability of a call increases. Blocking probability is the probability that call will be blocked while attempting to seize the circuit or it can be explained as the probability that the connection cannot be established due to insufficient transmission resources in the network. So an accurate estimation of traffic helps to minimize the blocking probability of a call. This paper is organized as follows .In section II paper present existing methods of traffic estimation. Section III present classification of traffic pattern based on traffic statistics. In section IV concept of adaptive extended Kalman filter estimator is explained. In section V the experiments are performed with the real time traffic dataset of mobile network of two service providers and results of simulation are discussed. Sections VI present the comparison of Adaptive EKF method with the Holt-Winters's Exponential Smoothing method [6] for traffic estimation of mobile network. Section VII concludes the paper with a summary.

II. Literature Review

In this section existing attempts to estimate the traffic data is discussed. Computation of blocking probability by the Erlang loss formulae is discussed in reference [2].The reference paper [3] introduces Auto Regressive Integrated Moving Average (ARIMA) model. This intends to develop statistical model allowing to estimate traffic through time series modeling. Problem associated with this model is that it is complex and time consuming process. And sufficient historical data needed to perform effective modal identification. In reference paper [4] Accumulation Predicting Model (APM) is introduced which improves shortcomings of ARIMA model. But this model requires large amount of data for estimation. Time series based prediction using spectral analysis is given in reference paper [5]. This method developed a sequential least square prediction. This represents real traffic flow well but Prediction accuracy decreases as forecasting horizon increases. Reference paper [6] relies on analysis of traffic data on cell. Forecasting technique shown in paper [6] is a quantitative forecasting method that uses mathematical recursive function to predict trend behavior. This uses time series modal to make prediction assuming that future will follow same pattern as past. But problem associated with the modal is that whenever traffic data set are less, than errors increases. Reference paper [8] present

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Extended Kalman Filter (EKF) for congestion avoidance in the road network .In reference paper [9] adaptive Kalman filter for real time measurement of available bandwidth is proposed. In reference paper [10] traffic density estimation using data fusion technique is introduced. For road traffic data, estimation with Kalman filter is performed .Advantage of the filter is, it's simple and flexible structure and prediction accuracy. When the measurements are nonlinear in nature Extended Kalman filter has to be used for estimating the traffic density which is given in reference paper [7].So in this paper an approach of traffic load estimation of different mobile service providers using adaptive EKF method is introduced.

III. ANALYSIS AND CLASSIFICATION OF TRAFFIC DATA

Traffic is a core indicator for measuring the load and utility rate of telecommunication network equipments. After analyzing traffic statistics it is clear that the traffic data shows different behavior of traffic in a cell. Three main factors are considered for classification of traffic behavior.

- *Day variations*. If we observe dataset during day time cell has the busiest hours, but in night time it has small value of traffic.
- *Week variation*. Traffic values will be different in week days than weekends or vice versa.
- *Accidental variation*. Holidays, national festivals etc. these variations are so abrupt and cannot be predicted by statistical analysis.

In the dataset day and week variation shows the periodicity or systematic cyclic behavior and may be predicted by statistical analysis .Actual traffic consist of all these kind of variations but algorithm applied for prediction will be complex and will be showing large error so classifying traffic behavior first according to above shown ways and then applying estimation technique for prediction of traffic will be easy and more accurate.

Traffic dataset of different cells of Bharat Sanchar Nigam Limited (B.S.N.L) Jabalpur provides the information of no. of traffic channel (TCH) and no. of transreciever equipment (TRE) associated with different cells. By using the available information first we have classified cells into three different categories and then we calculated the acceptable limit of traffic from the erlang loss formulae given in [2]. It means that if traffic goes beyond this value so there will be a requirement of capacity planning or allocation of more no. of channels in order to reduce the blocking probability. Now the classification of cells based on no. of TRE's and TCH per cell is shown below.

- *High intensity traffic cell*. The cell where the traffic intensity data was maximum and full no. of TCH &TRE's are allotted is considered as high intensity traffic cell. Example of these cells are given below in table1 .In cell1 and cell2, 4 TRE's were allotted. Each TRE can accommodate 8 no. of channels so maximum 32 channels can be accommodated by 4 TRE'S but 3 channels are reserved for broadcast and for synchronization purpose, so maximum 29 channels are allotted. And all are utilized by cell 1 and 2 which is shown in table 1.Whereas in cell 3, TRE's allotted were 8 and all 58 channels are utilized, example of these cells can be the market area, railway station etc.

TABLE1. ANALYSIS OF HIGH INTENSITY TRAFFIC CELL ON 14TH MAY 2013

cell	No. of TCH per cell	No. of TRE's	TCH load (Erlang) is calculated at 2% blocking [2]
Cell1	29	4	21.0
Cell2	29	4	21.0
Cell3	58	8	47.8

- *Medium intensity traffic cell*. The cell in which traffic values are comparatively of lower intensity and no. of TCH per cell are less than the earlier case, but these cells cannot be neglected because average traffic has large values during some hour in a day or several day in a week. Example of this cell can be residential areas. Below shown in table 2 example of cells. In these cells no .of TRE's are same but no of TCH per cell is less than what it can accommodate .so if there is a condition when TCH load increases so to maintain voice quality more no of channels can be allotted.

TABLE2.ANALYSIS OF MEDIUM INTENSITY TRAFFIC CELL ON 14TH MAY 2013

Cell	No. of TCH per cell	No. of TRE's	TCH load (Erlang) is calculated at 2% blocking [2]
Cell4	27	4	19.3
Cell5	25	4	17.5
Cell6	26	4	18.4

- *Low intensity traffic cell.* The cell where traffic values are least of the available dataset. The example of these cells can be rural areas. Table 3 shows example of the low intensity traffic cell where only 2 TRE's are allotted and no. of TCH per cell is less.

TABLE3.ANALYSIS OF LOW INTENSITY TRAFFIC CELL ON 14TH MAY 2013

cell	No. of TCH per cell	No. of TRE's	TCH load (Erlang) is calculated at 2% blocking [2]
Cell7	11	2	5.84

We have classified cells into three different categories above. As the traffic in a particular cell increase, then to reduce

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

congestion in the particular cell effective capacity planning is required. But effective capacity planning requires accurate the traffic estimation, now in the next section of the paper the concept extended Kalman filters for estimation of mobile traffic is described.

IV.EXTENDED KALMAN FILTER

The Kalman filter is a recursive filter that estimate the state of linear modal based on the last estimate of state and a no. of normally distributed noisy observation [7]-[11]. When made applicable to nonlinear model Extended Kalman Filter (EKF) can be used where the linearization of the nonlinear model around its current state is used. Over the last two decades EKF has been successively applied for traffic state estimation which is shown in reference paper [8]-[10].Traffic state at time k is uniquely described by the vector x_k of the cell .The EKF is based on nonlinear state space equation [1] which, in this case expresses the density vector as a function of density in the previous time in step plus process noise $\mathrm{Q_{k-1}}$.

$$x_k = f(x_{k-1}) + \mathrm{Q_{k-1}} \quad (2)$$

$f(x_{k-1})$ Denotes the nonlinear state transition matrix.EKF further more uses measurement equation describing the measurement vector as a function of x_k and with the measurement noise r_k.

$$y_k = h(x_k) + r_k \quad (3)$$

Function $h(x_k)$ expresses the function that maps the density to a variable in same dimension as the measurement.y_k denotes the vector of all the measurement, r_k is measurement noise

EKF algorithm consists of two steps i.e. prediction step and a correction step , the model under consideration is used to predict a new state vector along with the error variance-covariance matrix The prediction step is defined by

$$x_k^- = f(x_{k-1}) \quad (4)$$

$$P_k^- = A_k P_{k-1} A_k^T + Q_k \quad (5)$$

P_k^- is a priori estimate of error covariance matrix of state vector finally matrix A_K is used for linearization of the model .

$$A_K = \nabla \frac{f(x_{k-1})}{x_{k-1}} \quad (6)$$

Note that nonzero derivative exist between two adjacent columns.

In the second step i.e. correction step, measurement used to make correction to the state. For the EKF state measurement also need to linearise around the current state. For this define H_k to be the derivative of the measurement mapping function to the state.

$$H_k = \nabla \frac{h(x_k)}{x_k} \quad (7)$$

Second step of EKF is given by

$$K_{k=} P_k^- H_k^T (H_k P_k^- H_k^T + r_K)^{-1} \quad (8)$$

$$x_k^+ = x_k^- + K_k\big(y_k - h(x_k^-)\big) \quad (9)$$

$$P_K = (I - K_k H_k) P_k^- \quad (10)$$

Where I is an identity matrix, and K_k is called the Kalman gain, which indicates how much the state, should be corrected .The result of EKF is a posteriori state vector x_k^+ which is a balanced estimate of traffic state given both the estimate of the model and the measurements. .x_k^- Is the prior value of states. $h(x_k^-)$ denotes the nonlinear measurement matrix

For estimating the traffic in the mobile network the first step is to store the initial observation of states, so first the function "hfun1" is called by using matlab function "feval", which store the initial states of the system. Then estimation of state by applying EKF is performed, in order to do so, first of all mean is predicted and then calculating the jacobian matrix for calculation of covariance, and then by utilizing the equations of EKF shown above final stage of estimation is performed. After calculating all the estimated values, the error is calculated between the true state and EKF estimated state of traffic intensity and then results of the simulation are shown in the next section.

V.EXPERIMENTAL RESULTS

Traffic dataset are collected from BSNL and IDEA CELLULAR, and then experiments are performed. Statistics consist of GSM traffic data of different cell of BSNL and MSC traffic data of IDEA CELLULAR. Hourly based statistics are taken, and then estimation with Adaptive EKF method is performed. Two experiments are performed with the available data. In experiment no.1 traffic estimation of cell 1 of BSNL JABALPUR is performed and in experiment 2 traffic estimation of IDEA CELLULAR NOIDA is performed. In the experiments historical traffic data is taken for analysis purpose, and based on those data sets Traffic estimation is performed .we have taken the traffic dataset of 144 hour to estimate the traffic of next 12 hours.

Experiment.1. First experiment is performed with the data collected from BSNL office Jabalpur. Simulation is performed in the matlab software and the result of simulation is shown below.

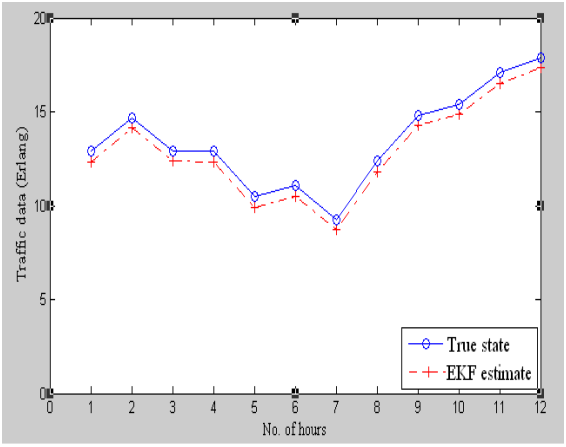


Fig. 1.Performance of traffic load estimation of 1 cell of BSNL JABALPUR on 16th may 2013

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fig. 1 shows the performance of EKF estimating the traffic load of 12 hours i.e. from 10.00 hours to 22.00 hours. Which includes busy hours also i.e. from 10.00 hour to 11.00 hour and from 7.00 hour to 8.00 hour of a cell associated with BSNL JABALPUR, and hourly based statistics is considered.

We estimated the accuracy of proposed method by Normalized Root Mean Square Error (NRMSE) which is calculated by following formula.

$$NRMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_k - \hat{x_k}}{x_k}\right)^2} \quad (11)$$

Where x_k is the real value of traffic at time k, and $\hat{x_k}$ is the predicted value of traffic at time k. n is the total no. of fitted forecast values.

For experiment 1 we have calculated the NRMSE by the formula shown above in equation 11, NRMSE of 0.0063 is calculated.

Now the accuracy of the proposed method for traffic estimation with respect to no. of iteration is shown below in fig. 2.

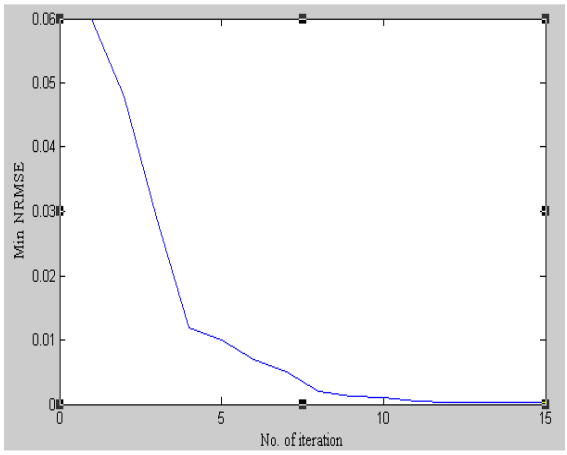


Fig. 2. No. of iteration v/s minimum NRMSE plot

Fig.2 shows the plot between NRMSE and no. of iteration and it is clear from the plot that as we increase no. of iteration NRMSE decreases and better accuracy in the result of estimation is found. The best result of NRMSE is 0.0002 which is shown in fig.2.

Now the parameter based analysis is shown next where first we vary the process noise by keeping the measurement noise constant and then measurement noise is varied by keeping the process noise constant and then its effect on estimation is described.

Parameter Based Analysis

1. *Vary Process Noise*. Now we first increase process noise Q from 0.01 to 0.1 and keep the measurement noise constant to 0.1 and then experiment for estimation of mobile traffic of 1 cell of BSNL Jabalpur is performed and then result of simulation is shown below in fig. 3.Minimum NRMSE of 0.0002 is calculated.

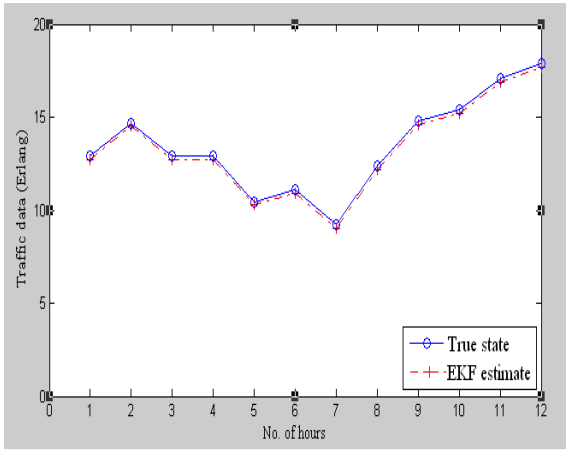


Fig. 3. Performance of traffic load estimation of 1 cell of BSNL JABALPUR on 16th may 2013

2. *Vary Measurement Noise*. Now we increase measurement noise r from 0.1 to 1 and keep the process noise constant to Q=0.01 and results of simulation is shown below in fig.4.Minimum NRMSE of 0.048 is calculated.

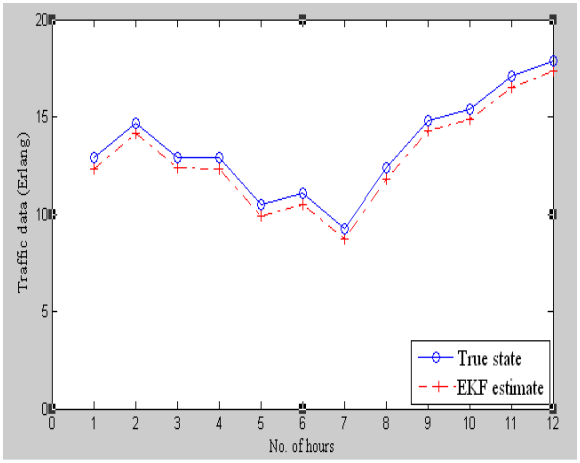


Fig.4.Performance of traffic load estimation of 1 cell of BSNL JABALPUR on 16th may 2013

In both the case of parameter based analysis Adaptive EKF method performed satisfactory for estimating the 12 hour traffic of a cell associated with the BSNL JABALPUR.

Experiment.2. In this experiment estimation is performed by considering the traffic data sets of IDEA CELLULAR, and then result of simulation is shown in fig. 5.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

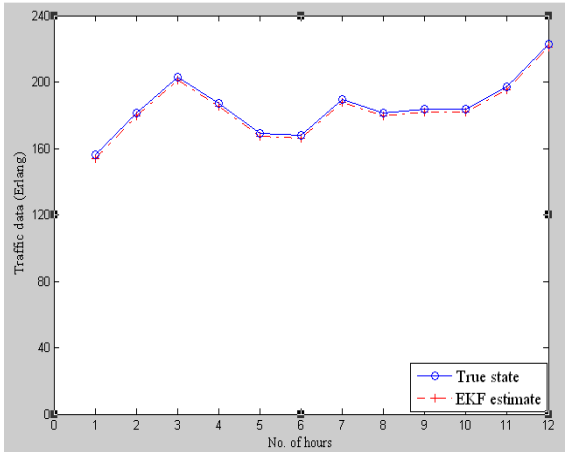


Fig. 5.Performance of traffic load estimation of IDEA CELLULAR NOIDA on 16th may 2013

Fig. 5 shows the performance of EKF estimating the traffic load of 12 hours i.e. from 10.00 hours to 22.00 hours. The NRMSE is then calculated by the by the formulae shown in equation 11, minimum NRMSE of 0.0003 is calculated.

Plot between NRMSE and number of iterations is shown in fig. 6.

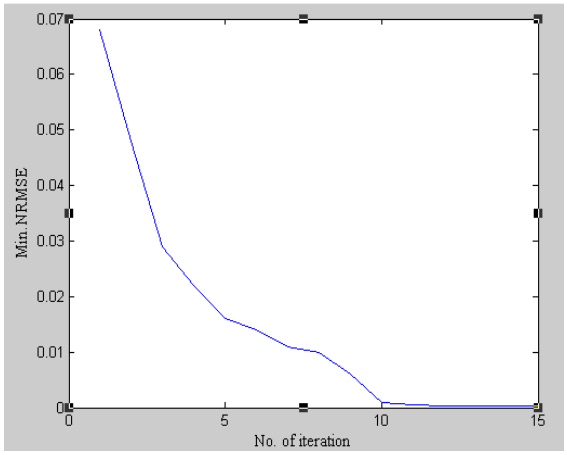


Fig. 6. No. of iteration v/s minimum NRMSE plot

It is clear from the plot shown above in fig.6 that as we increase the no. of iteration then the accuracy of the proposed method for estimation increases. The best result of NRMSE is 0.0003 which is shown in fig.6.Now the parameter based analysis is shown below.

Parameter Based Analysis

1. *Vary Process Noise*. Now we first increase process noise from 0.01 to 0.1 and keep the measurement noise constant to 0.1 and result of simulation is shown below in fig.7.Minimum NRMSE of 0.0010 is calculated.

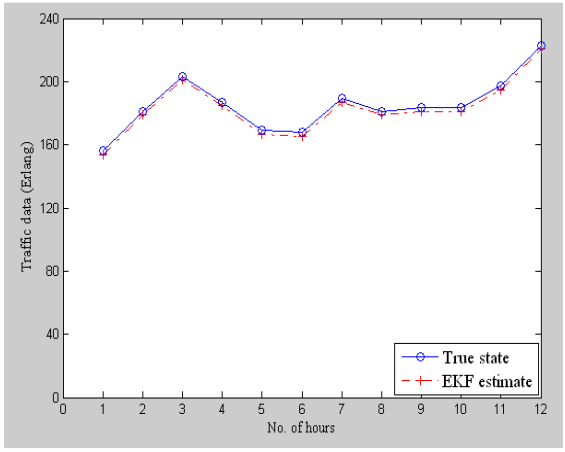


Fig.7.Performance of traffic load estimation of IDEA CELLULAR NOIDA on 16th may 2013

2. *Vary Measurement Noise*. Now we increase measurement noise from 0.1 to 1 keeping the process noise constant to Q=0.01 and results of simulation is shown below in fig. 8. Minimum NRMSE of 0.012 is calculated.

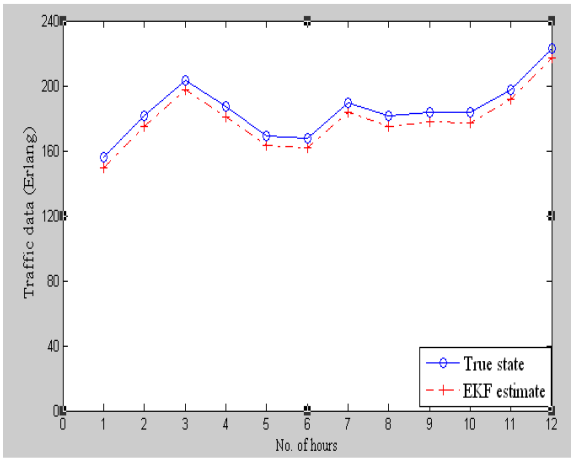


Fig.8.Performance of traffic load estimation of IDEA CELLULAR NOIDA on 16th may 2013

It is clear from the fig.7 and fig.8 that by varying process noise and measurement noise Adaptive EKF method performed accurately.

Now comparison of two methods for mobile traffic estimation is given in the next section.

VI COMPARISON OF ESTIMATION METHODS

TABLE 4: PERFORMANCE OF ADAPTIVE EKF METHOD AND HOLT-WINTER'S EXPONENTIAL SMOOTHING [6] FOR MOBILE TRAFFIC ESTIMATION

Method	Minimum NRMSE
Holt-Winters's Exponential Smoothing [6]	0.1228
Adaptive EKF	0.0063

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 4 shows the performance of two methods for estimation of mobile traffic. From the different techniques discussed in literature review section Holt-Winters's Exponential Smoothing method [6] particularly estimated the traffic of mobile network. So in this section a comparison of Adaptive EKF method is being made with Holt-Winters's Exponential Smoothing. While estimating the traffic of mobile network Holt-Winters's Exponential Smoothing method gives the NRMSE of 0.1228 (in the case when the traffic intensity is upto 20 erlang). And this method [6] has considered the whole 24 hour data to be the data estimated. Whereas Adaptive EKF method presented in this paper for traffic estimation gives the NRMSE of 0.0063(for the same range of traffic).But in our estimation process we have considered the 12 hour data i.e. from 10.00 hours to 22.00 hours, (which includes busy hours also i.e. from 10.00 hour to 11.00 hour and from 7.00 hour to 8.00 hour), to be the data estimated. So this has made the analysis easier and more accurate results are found.

VII.CONCLUSION

In this paper traffic intensity estimation of mobile communication network using Adaptive Extended Kalman Filter method has been successfully implemented. Real time Traffic intensity datasets are collected from BSNL and IDEA CELLULAR. These datasets are pre-processed and then made compatible for estimation model. Adaptive EKF method successfully estimated the traffic based on historical data. And then parameter based analysis by varying process noise, measurement noise, no .of iterations being done. Results of simulation shows performance quiet efficient. Compared with Holt winters method [6] for estimation of traffic intensity which gives the NRMSE of 0.1228(when traffic intensity is upto 20 erlang),estimation with Adaptive EKF method gives the NRMSE of 0.0063 (for the same range of traffic).And by increasing no. of iteration minimum NRMSE of 0.0002 is calculated.

ACKNOWLEDGMENT

We acknowledge two mobile service providers BSNL and IDEA network for providing real time traffic data. This work is partially supported by Madhya Pradesh Council of Science and Technology (M.P.C.S.T.) BHOPAL for research work.

REFERENCES

[1] Simon haykin "Adaptive filter theory" 3rd edition, Prentice Hall.

[2] William C.Y. Lee Mobile Cellular Telecommunication, 2nd edition McGraw-HILL international edition.

[3] Cesar Augusto Hernandez Suarez , Octavio Salcedo Parra, Luis Fernando "Traffic model based on time series to forecast traffic future values within a Wi-Fi data network,"4th International Conference on WiCOM 2008, pp.1–4.

[4] Yanhua Yu, Meina Song, Zhijun Ren, lunde Song "Network Traffic Analysis and Prediction Based on APM," 6th International Conference on Pervasive Computing and Applications (ICPCA), 2011,pp.275-280

[5] Tigran T. Tchrakian, Biswajit Basu,Member, IEEE, and Margaret O'Mahony "Real-time traffic flow forecasting using spectral analysis," IEEE transactions on Intelligent Transportation Systems vol. 13, no. 2,june 2012,pp.519-526.

[6] Denis Tikunov,Toshikazu,Nishimura "Traffic Prediction For Mobile Network Using Holt-Winters's Exponential Smoothing,"15th International Conference on Software, Telecommunications and Computer Networks, SoftCOM 2007, pp.1-5. .

[7] G.Bishop and G.Welch "An Introduction to Kalman Filter," SIGGRAPH 2001, Course 8, 2001.

[8] Sung-Soo Kim, Yong-Bin Kang "congestion avoidance algorithm using Extended Kalman filter," Proceedings of the 2007 International Conference on Convergence Information Technology (ICCIT '07), pp.913-918.

[9] Svante Ekelin, Marlin Nilsson,Erik Hartikainen,Andreas Jhonso,.Jan-Erik Mangs,Bob Melander,and Mats Bjorkman "Real-Time Measurement Bandwidth using Kalman Filtering," Network Operations and Management Symposium, 2006. NOMS 2006. 10th IEEE/IFIP, pp.73-84.

[10] R. Asha Anand, Lelitha Vanajakshi, and Shankar C. Subramanian"Traffic Density Estimation under Heterogeneous Traffic Conditions using Data Fusion," IEEE Intelligent Vehicles Symposium (IV)Baden-Baden, Germany, June 5-9, 2011,pp.31-36.

[11] Iwao Okutani and Yorgos J. Stephanedes. "Dynamic prediction of traffic volume through Kalman filtering theory," Transportation Research Part B: Methodological, Volume 18, Issue 1, February 1984, pp 1–11.

Test Framework Development Based Upon DO-278A Using ISTQB Framework

Raha Ashrafi*
Computer Engineering Department
North Tehran Branch , Islamic Azad University
Tehran , Iran

Ramin Nassiri PhD
Computer Engineering Department
Central Tehran Branch, Islamic Azad University
Tehran, Iran

***Abstract*—DO-278A is an FAA Standard which was published for CNS/ATM software development in 2011 while ISTQB[1] is the worldwide software testing standard framework. Software which are developed according to DO-289A are verified by Review, Analysis and a few other methods but they are not actually sufficient for testing so it would be quite reasonable to deploy ISTQB techniques for that purpose. This paper intends to show that ISTQB techniques may successfully work for DO-278A verification.**

Keywords: component; Software Aviation system; CNS/ATM; DO-278A;, ISTQB; Verification.

I. INTRODUCTION

Software failures are common. In most cases, although these failures can cause discomfort and inconvenience, but they do not cause any long-term or serious risk, however, in some systems, failures could result in large economic losses, physical damage or threats to human life. This system is called Critical System. Critical systems fall into three categories: [1]

- Safety Critical system
  Failure of this system will result in loss of human life or serious damage to environment.
- Critical Mission system
  Failure of this system will result in purposeful activities.
- Business Critical system
  Failure of this system will result in high economic losses for business users.

The high cost of failure in Critical system implies that they need reliable methods and unique techniques for development and more than 50% of total development cost is spent on verification.

Development techniques of critical system include: [1]

- Fault avoidance
  This technique is used to reduce the likelihood of defects into the system
- Fault detection and elimination
  Verification and Validation techniques are used to detect defects and eliminate them
- Fault tolerance
  These techniques are used to ensure that errors do not cause system failure.

Testing is a common method for checking the fulfillment of customer specifications and requirements. However testing is one of V&V techniques. Along with and after the development process, the developed product shall be investigated to meet customer requirements and to be checked for errors and defects.

To meet this purpose, various tests are executed on the software before they become operational. There are several standards for testing. One of them is ISTQB as an international software test framework.

CNS/ATM software is a Safety Critical software which uses V&V techniques for development and there exists a special Standard for their development.

FAA published DO-178 as a guidance for airborne software development at 1982 and published DO-178B as a modified version of DO-178 at 1992 and for the first time published DO-278 as a supplement for development of CNS/ATM software that is used with dependency to DO-178B. DO-278 is a dependent document that is used concurrently with DO-178B for development of CNS/ATM software. From 1992 according to questions raised and modifications needed, FAA published DO-178C and DO-278A while these are the latest versions for software producers. DO-278A is a standalone document for developing CNS/ATM software that is used for Air Traffic control purposes.

This paper compares DO-278 and FAA standards for producing ground-based software of CNS/ATM, and ISTQB, standard for Software testing to find a framework testing according DO-278 and ISTQB.

* Corresponding author

[1] International Software Testing Qualification Board

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

II. DO-278A

A. Assurance Level

DO-278A that was published in 2011 is a stand alone standard for producing CNS/ATM ground-based software. However, its previous version, DO-278, is used together with DO-178B which is a standard for airborne software. According to DO-178B, the failure condition caused by software errors is divided into five categories. (Table 1)

Table 1 Failure Condition [2]

Severity	Description
Catastrophic	Its cost isn't compensable such as human lives
Hazardous	The situation is difficult to control and can lead to a catastrophe if it gets out of control
Major	The situation is under control but failing to control the situation becomes hazardous
Minor	The situation is under control
No Safety Effect	The situation that would not effect safety

According to five levels of failure condition, software assurance levels are defined. It shows that if the assurance level is A, the catastrophic incidents will not happen or in order to avoid catastrophic incidents, the software should be at level A of assurance. (Table 2)

Table 2 Assurance Level [3]

Severity	Assurance Level
Catastrophic	A
Hazardous	B
Major	C
Minor	D
No Safety Effect	E

Table 3 associates DO-278A with DO-178C. AL4 in DO-278A doesn't have any equivalence in DO-178C. It was designed for ground-based system when AL3 is strict and AL5 is too permissive for that. [4]

Table 3 Assurance Level association [2]

DO-278A	DO-178C
AL1	A
AL2	B
AL3	C
AL4	Not Equivalent
AL5	D

B. Software Life cycle processes [2]

Software life cycle processes in DO-278A are:

- Software planning processes;
- Software development processes;
- Integral processes.

Software development processes are:

- Software requirements process;
- Software design process;
- Software coding process;
- Integration process.

Integral processes are:

- Software verification process;
- Software configuration management process;
- Software quality assurance;
- Approval liaison process.

The integral processes are performed concurrently with software planning and development processes throughout the software life cycle according Fig 1. In this paper, the review of verification processes is considered.

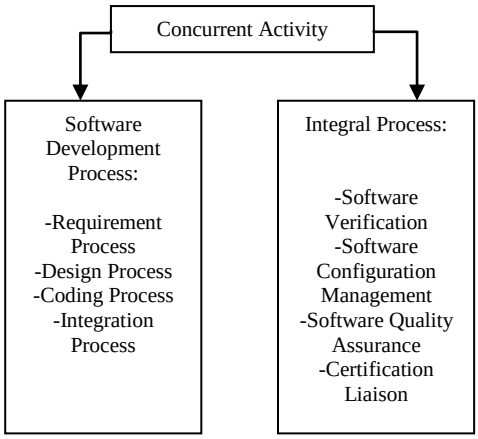


Figure 1 Concurrent Activity

III. ISTQB FRAMEWORK [5]

ISTQB is the software-testing standard, which is published in 2002. In this standard, techniques used are divided into the general classes:

- Static
- Dynamic

A. Static techniques

Static category detects error without running the software and includes two main techniques:

- Review
- Static analysis

In which, each of them have their sub-branches as follows:

- Review:
  - Informal review
  - Walk through
  - Technical review
  - Inspection
- Static analysis:
  - Data flow
  - Control flow

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. Dynamic techniques

Dynamic technique has three main categories in which each of them has its subclasses as follow:

- White box technique
  - Statement Coverage
  - Decision Coverage
  - Loop Coverage
  - Condition Coverage
  - Multiple Conditions
  - Basic Path
- Black box technique
  - Equivalence Partitioning
  - Boundary Value Analysis
  - Decision Table
  - State Transition
  - Comparing
- Experience based technique
  - Error Guessing
  - Exploratory Testing

IV. Review Consideration of DO-278A and ISTQB

Generally, DO-278A verification is done by review, analysis and testing. However, there is not any special technique for verification.

Subject to verification objectives which is described in tables of annex A of DO-278A, ISTQB techniques can be used for DO-278A verification.

A. Static Techniques

Correspondence table of DO-278A objectives and static ISTQB techniques are described in Table 4. According to objective tables, the following result is derived:

From among ISTQB Static techniques, informal review is only used for coding of DO-278A. Since DO-278A is about Safety-Critical software and informal review is not used for such software, then the results of ISTQB Static Review that can be used in DO-278A are:

- Peer Programming (Informal Review)
- Walkthrough
- Technical Review
- Inspection
- Data Flow
- Control Flow

B. Dynamic Techniques

Correspondence between DO-278A objectives and Dynamic ISTQB techniques are described in Table 5. According to objective tables, following result is derived:

Testing strategy of DO-278A is based on requirement testing, so it must be considered in selecting the test cases and also requirement coverage must be obtained in all assurance levels.

From among Dynamic technique of ISTQB all of them are used except:

- Decision Coverage;
- Basic Path;
- Decision Table;
- Comparing;

And other techniques are used to Verification under DO-278A.

Table 4 Review and Analysis f DO-278A and Static techniques of ISTQB

ISTQB / DO-278A	Review				Static Analysis	
	Informal Review	**Walkthrough**	**Technical Review**	**Inspection**	**Data Flow**	**Control Flow**
Review and Analysis of High Level Requirement		✓	✓	✓		
Review and Analysis of Low Level Requirement		✓	✓	✓		
Review and Analysis of Software Architecture		✓	✓	✓	✓	✓
Review and Analysis of Source Code	✓	✓	✓	✓	✓	✓
Review and Analysis of Integration			✓	✓	✓	✓

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 5 Testing of DO-278A and Dynamic Technique of ISTQB

ISTQB / DO-278A	White Box						Black Box				
	Statement Coverage	**Decision Coverage**	**Loop Coverage**	**Condition Coverage**	**MC/DC**	**Basic Path**	**Equivalent Partitioning**	**Boundary Value**	**Decision Table**	**State Transition**	**Comparing**
Normal Test case			✓				✓	✓		✓	
Robustness Test case	✓		✓				✓	✓		✓	
Test Coverage of Software Structure	✓			✓	✓						

V. CONCLUSION

DO-278A is for producing CNS/ATM software and ISTQB is a software testing Framework.

By using ISTQB techniques in DO-278A we could be able to achieve a robust testing technique to fulfill testing requirements. ISTQB Techniques deployed for CNS/ATM Software are:

- Static
  - Informal review
  - Walk through
  - Technical review
  - Inspection
  - Data flow
  - Control flow
- Dynamic
  - Statement Coverage
  - Loop Coverage
  - Condition Coverage
  - Multiple Conditions
  - Equivalence Partitioning
  - Boundary Value Analysis
  - State Transition

This paper aimed to show the capability of ISTQB Framework to test mission-critical systems as a proof of the compatibility of that framework with such systems.

REFERENCES

[1] I. Sommerville, "Software Engineering (Eighth Edition) " , Person , 2006

[2] DO-278A “Software Integrity Assurance Considerations for Communication , Navigation , Surveillance and Air Traffic Management (CNS/ATM) systems” , SC-205 RTCA Inc. , 2011

[3] DO-178B “Software Considerations in Airborne Systems and Equipment Certification” , RTCA Inc. , 1992

[4] L. Rierson, “Development safety critical software” , CRC Press , 57 (2013)

[5] D. Graham, E. Veenendaal, I. Evans and R. Black, “ Foundations of Software Testing” , ISTQB , 2007

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Active Queuing Mechanism for WCDMA Network

Vandana Khare [1], Dr. Y. Madhavee Latha [2], Dr. D. SrinivasRao [3]

[1] Associate professor & HOD ECE, RITW, Hyderabad, India

[2] Professors & Principal MRECW, Secunderabad, India

[3] Professors & Head ECE Department, JNTU, Hyderabad, India

Abstract: The network traffic in the upcoming wireless network is expected to be extremely non-stationary and next generation wireless networks including 3rd generation are expected to provide a wide range of multimedia services with different QoS constraints on mobile communication because of that there is no guarantee for a given system to provide good quality of service. So that there is a need to design an efficient queuing mechanism by which the QoS is going to be improved for wideband services. This paper proposes an efficient active queue management mechanism to control the congestion at the router. This paper is mainly aimed towards the WCDMA scheme for wideband services like video. So that it integrates the WCDMA with IR-RAN using active queue management. By simulation result our proposed WCDMA architecture along with active queue management (AQM) will achieves the effective peak signal to noise ratio (PSNR).

Keywords: WCDMA, IP-RAN, QoS, ,PSNR.

I.INTRODUCTION

The consumers of the future will put new requirements and demands on services. However, the fact is that today we already use different kinds of multimedia services, i.e. services that to some extent combine pictures, motion and audio. These include TV, video and the Internet. Many of these applications have become fundamental elements of our lives because they fulfill basic needs. So to access all these services there should be minimum loss during data transmission. This can be achieved by designing an efficient medium having minimum losses. Now a days the various types of services accessing by multi users at a time contains an important data. So this important data has to be received in the way such that there should be maximum perceived quality. This both features can be obtained effectively by integrating the WCDMA communications with the IP networks. The main aim in these IP-RAN networks is maintaining the good quality of service as well as at a time it has to achieve multi high data rate services. The aim of multi high rate service accessing can be achieved by WCDMA [1], because it comes under wideband communications. Another aim of achieving good quality of service can be achieved by designing an efficient congestion [3], [4] control method at the router. With the increase in the data access using these protocol, demands for larger bandwidth in coming future. Increasing bandwidth may not be a suitable solution as it is economically non-advisable. The decrease in the resources may lead to congestion [4] in the network resulting to complete collapsing of the network. A mechanism is hence required to overcome these problems so as to support larger data in the constraint resource to provide fair routing with least congestion.

This work proposes an active queue management for control of congestion in IP-RAN network which is utilizing the Wide band code division multiple accesses (WCDMA)[2] at the transmitter and receiver stages, to maximize network capacity while maintaining good voice quality. The principle underlying the scheme is regulation of the IP RAN load by adjusting the queue according to the priority. IP routers using a drop tail mechanism during congestion could produce high delays and bursty losses resulting in poor service quality. Use of active queue management at the routers reduces delays and loss correlation, thereby improving service quality during congestion. This project objective is to implement a efficient congestion control mechanism on router using active queue management improving the performance of IP-Based Radio Access Network.

The rest of the paper is organized as follows: section II gives the details about the system model used. The complete detail about the design of an active queue management for congestion avoidance at the router is illustrated in section III. Next, section IV gives the results evaluation for the proposed approach, finally conclusions are provided in sectionV.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

II.SYSTEM MODEL

This section gives the illustration about the basic system model used to implement the proposed approach. The proposed approach utilizes the WCDMA system model to provide the multiple accessing because in wideband systems, the transmission bandwidth of a single channel is much larger than the coherence bandwidth of the channel. Code division multiple access technique comes under wideband systems. The general communication architecture of the proposed approach is shown below.

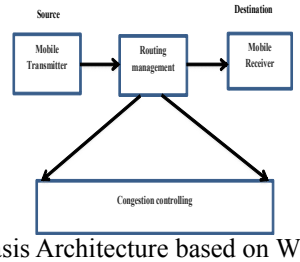


Fig.1.Basis Architecture based on WCDMA

The basic communication system modeling utilizing the WCDMA as multiple accessing techniques shown in fig.1 is illustrated as follows:

A. **Transmitter design**

In Wide band Code Division Multiple Access (WCDMA) system the input band data signal is added to a high rate-spreading signal. This spreading signal is formed from a pseudo-noise code sequence, which is then multiplied by a Walsh code for maximum orthogonality to (i.e. to have low cross-correlation with) the other codes in use in that cell. Typically, CDMA pseudo-noise sequences are very long, thereby giving excellent cross correlation characteristics. The IS-95 system can be thought of as having many layers of protection against interference. It allows many users to co-exist, with minimal mutual interference. The spreading signals from all the users are summed up and transmitted through the DPSK modulator. The following fig.2 gives the basic design of the transmitter using in this approach.

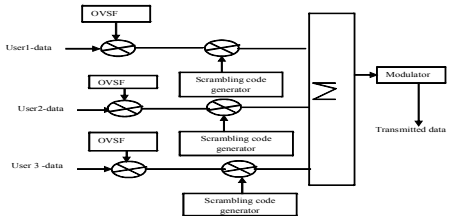


Fig.2. Transmitter block diagram

Transmissions from a single source are separated by channelization codes, i.e., download connections within one sector and the dedicated physical channel on the uplink. The OVSF channelization code preserves the orthogonality between different physical channels using a tree-structured orthogonal code. Next the Scrambling codes make the direct sequence CDMA (DS-CDMA) technique more effective in a multipath environment. It significantly reduces the auto-correlation between different time delayed versions of a spreading code so that the different paths can be uniquely decoded by the receiver. Additionally, scrambling codes separate users and base station sectors from each other by allowing them to manage their own OVSF trees without coordinating amongst themselves. After that the spreader messages from multiple source nodes are summed together for transmissions are spectrally overlapped and time-overlapped. Adaptive power control schemes are employed in WCDMA technology for efficient transmission of messages. The differential encoding process at the transmitter input starts with an arbitrary first bit, serving as a reference, and therefore the differentially encoded sequence d_k is generated by using the logical equation

(1)

Where m_k is the input binary digit at time KT_b and d_{k-1} is the previous value of the differentially encoded digit. The symbol denotes module-two addition, and the use of an overbar denotes logical inversion. The differentially encoded sequence d_k thus generated is used to phase-shift key a carrier with the phase angles 0 and radians. The block diagram of the DPSK transmitter, which is used in this thesis, is shown in the fig above. It consists in part, of a logic network and a one-bit delay element interconnected so as to convert an input binary sequence m_k into a differentially encoded sequence d_k in accordance with the above equation.

B. Receiver Design

At the receiver, the inverse operation of transmitter is performed to recover the original signal, while interfering signals are considered as noise, and mainly suppressed by the matched filter. The Multi User Detector block in the receiver carries with the bank of matched filters to dispread the user data at the receiver output.

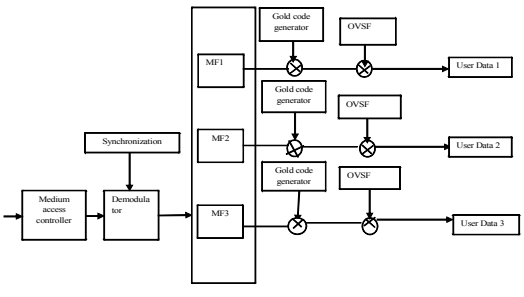


Fig.3.receiver Block diagram

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The basic function of the DPSK demodulator is illustrated here. At the receiver, a reference carrier is created from the input signal. This recovered carrier is at the same frequency as that of the original carrier, but without any phase changes, in the sense the reference carrier has a constant phase, the recovery scheme used is explained. The recovered carrier is mixed with the input modulated signal to retrieve the unmodulated signal. At the receiver input, the received DPSK signal plus noise is passed through a band-pass filter centered at the carrier frequency f_c, so as to limit the noise power. The filter output and delayed version of it, with the delay equal to the bit duration T_b, are applied to a correlator. The resulting correlator output is proportional to the cosign of the difference between the carrier phase angles in the two correlator inputs. The correlator output is finally compared with a threshold of a zero volts, and a decision is thereby made in favor of symbol 1 or symbol 0.This discrimination between the received codes is accomplished by means of a matched filter, whose output indicates when a particular code sequence is detected in the input data stream.A matched filter is a filter whose frequency response is designed to exactly match the frequency spectrum of the input signal. These filters are used as signal processors in communications receivers to calculate the correlation between the transmitted signal and the received signal.

In WCDMA systems the matched filter is tuned to match a code sequence, which is expected to be contained within the digital samples entering the system receiver. The matched filter indicates when this code sequence is detected in the input data stream. The output of a matched filter will be a score value. A higher score value indicates a more tuned match with the code sequence of the received data stream. This is also called correlation, and hence a high score value represents a good correlation of input with the code sequence of interest. Unlike a transmission filter where the continuous data stream entering the filter is modified to form a new continuous data stream, a matched filter output must be considered to be a flow of individual results. These results must then be analyzed to identify the individual point where the match occurred. The message retrieved from the match filters consists of message with the spreading code. Once the message of each user is differentiated using the match filter Bank the retrieved message bits are passed down to Dispread blocks, where these message bits are again multiplied by the spreading code developed by the gold code generator in the receiver system. The dispreading of the message takes place exactly opposite to the spreading method. Here each n-bit is represented by its equivalent one bit of message.

In this way a successful communication is going to be established between source and destination. An important challenge in this schemes addressing congestion occurrence at the channel (router) from entering the congestion state rather than having to recover once congestion has occurred. A scheme that allows the network to prevent congestion is called a router design for congestion avoidance is discussed in next section.

III. ROUTER DESIGN TO AVOID CONGESTION

This section gives the design of a router module which in present between transmitter and receiver. The main purpose to design this router is to control the congestion [3], [4] occurring at the router. There are so many congestion control techniques are proposed to avoid the congestion ta the router like drop tail router, random early detection, active queue management, etc., this section gives the complete description about the active queue management, because this was gaining more research interest in recent days.

A. ROUTER MECHANISIMS

We term a group of successive packets in a queue, having the same IP source-destination addresses and the same port numbers, as “burst of packets belonging to the same flow”.

The router active queue management algorithm detects *incipient congestion*. In this case, the packet to be sent is checked in order to identify two different conditions: 1, the packet is ECN marked, and 2, the packet is the last packet of a burst of packets belonging to the same flow and stored in the router’s queue. When the packet is not marked (condition 1 does not hold), the ECN bit is set in the IP header. If condition 2 applies and if the packets of the burst do not hold any ECN mark, an ICMP source quench message is returned to the source of the burst.

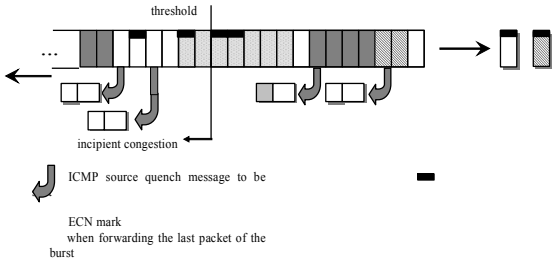


Fig.4. Routing mechanism

The ECN bit set in the IP header of a packet is used by the successive congested routers as an indication that the burst has already passed through at least one congested router. Hence, an ICMP source quench was already transmitted to the sender for this burst and an additional ICMP source quench transmission is avoided. When the router’s queue becomes full, the queue enters the *congestion* state and the arriving packets are dropped. The queued

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

packets to be sent are checked as in the incipient congestion case.

B.Sender mechanisms

The sender reaction to network congestion is the most sensitive part of a congestion avoidance scheme, influencing the loss rate at the bottleneck link (high if the sending rate is too high), the efficient utilization of resources (low if the sending rate is too slow), and the fairness among the flows involved. Negotiations between sender(s) and receiver(s) by means of end-to-end flow control mechanisms allow establishing data transfer rate boundaries agreed by receiver(s). These boundaries are hereinafter called limits. In the case of TCP[5], the flow control is ensured by a window mechanism that uses a field in the acknowledgments to advertise the buffer space of the receiver to the sender. In the case of UDP, flow control mechanisms, if any, are handled at a higher layer. We focus below on the algorithm used to adjust the data transfer rate offered by the sender to ICMP source quench. As long as no ICMP source quench has been received since the data transfer started, the sources keep on sending data at the rate monitored by the upper layer mechanisms (e.g. TCP for TCP flows, application for UDP flows). When the first ICMP source quench arrives indicating congestion or incipient congestion, additional mechanisms are activated at the sender side.

IV.RESULTS

This section gives the complete illustration about the performance evaluation of the proposed approach. There are 30 wireless nodes move randomly in a given 1000x1000 m^2 square. Video sequences are in the .avi format. We encoded the video into MPEG4 formatted file, and transmitted through the network shown below. Finally compare the file after transmission with the original file, and calculate the PSNR (Peak Signal-to-Noise Ratio) value. PSNR is one of the most widespread objective metrics to assess the application-level QoS of video transmissions. First to evaluate the numerical analysis we have to evaluate the mean square error between the original video and recovered video. Then the PSNR can be evaluated by applying the logarithm for that MSE. The formulae for MSE and PSNR evaluation are shown below.

MSE= (2)

Where the recovered video is sequence and is the original video sequence.

PSNR=10*log(MSE) (3)

The wireless bandwidth is set as 11Mb. Node communication radius is set as 300m. The length of interface queue is set 100.

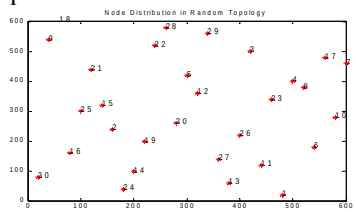


Fig5.Random topology in networks

In this work to evaluate the proposed approach we had considered a network having 30 nodes. The area of the network considered is 600 m^2. In this network the nodes are distributed in random fashion. The x-axis denotes length of the network along horizontal direction and the y-axis denotes the length along the vertical direction

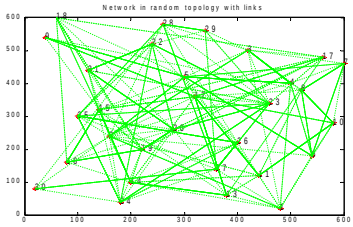


Fig6. Network in random topology with links.

The above figures shows the complete links provided from each and every node to each and every node. For this purpose the distance from node to node has to be evaluated. Then only the links are going to be plotted by comparing it with the network range. Thus to provide a link between two nodes the distance between that two nodes should be less than the network range.

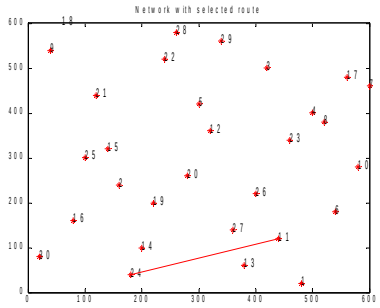


Fig7. Network with selected route .

The above figure shows the path established between two nodes. Here the node 21 is considered as source node and the node 11 is considered as destination node. To establish the path between these two nodes first the source node has to send the request packet to destination node then the destination node has to be acknowledge an acknowledgement to source. If the source receives

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

theacknowledge from the given destination node then the path is said to be established. Then only the communication for information transfer is going to be occurred between those two nodes.

Fig8. Original video sequence

The above figure represents the original video sequence taken to process through the selected path between source node and destination node. The given video file format is qcif.yuv. The proposed algorithms applied on this video one by one and the performance is analyzed by evaluating the PSNR for obtained video at destination

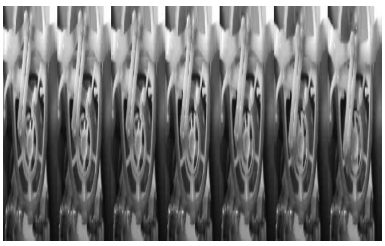


Fig9. Extracted frames

Generally a video file is considered as a group of images (frames). So to process any operation on an video in the first stage we have to extract the frames from it. The above figure denotes the extracted frames for the given video sequence in the fig11.

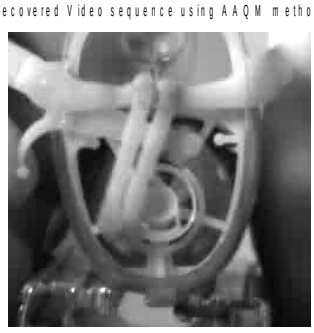


Fig10. Recovered video sequence using AQM

The above figure denotes the recovered video sequence at the decoder. It is almost same as is it is of the input video sample given at the source node. So that the proposed method is said to be efficient.

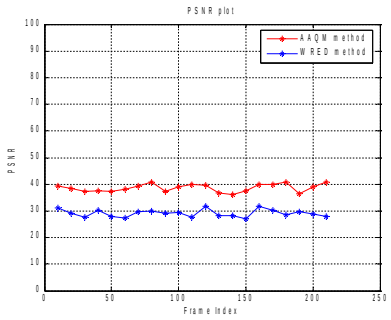


Fig11.PSNR plot

The above figure denotes the PSNR plot for the proposed approach and also for previous approach. First to evaluate the numerical analysis we have to evaluate the mean square error between the original video and recovered video using equation (2). Then the PSNR can be evaluated by applying the logarithm for that MSE. This is evaluated by equation (3). From the adobe figure.13 it is clear that the obtained PSNR value is high for proposed approach for a given frame index.

V.CONCLUSION

This paper introduces a new approach using WCDMA Network for fairness by means of the router round trip time concept used by our mechanism. We evaluated the mechanism behavior by simulations and our results show that it achieves fair sharing of network resources using little bandwidth control overhead. Due to space limitations, many details of the algorithm and simulation results are not shown in this paper.

REFERENCES

[1] VenugopalV.V and Ashok Mantravadi, ”Thecoding spreading tradeoff in WCDMA systems”, IEEE Select .Areas Commun.,vol.20,pp.396-408,Feb2002.

[2] Mingxi Fan and Kai-yeungSiu,”A dual-mode multi-user detector for DS-CDMA systems”, IEEEJ.select. Areas Commun, vol.20, pp.303-309, Feb2002.

[3] R. Rejaie, M. Handley, D. Estrin. RAP: An End-to-end Rate-based Congestion Control Mechanism for Realtime Streams in the Internet. *Proceedings of IEEE INFOCOM’99*, March 1999.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[4] D De Lucia, K. Obraczka. A Congestion Control Mechanism for Reliable Multicast. IRTF RMRG meeting, September 1997.
[5] J. Padhyle, J. Kurose, D. Towsley, R. Koodli. A TCP-friendly Rate Adjustment Protocol for Continuous Media Flows over Best Effort Network. UMASS CMPSCI Technical Report 98-47, October 1998. An Extended Abstract will appear in the *Proceedings of SIGMETRCICS'99*.
[6] D. Sisalem, H. Schulzrinne. The Loss-Delay Based Adjustment Algorithm: A TCP-Friendly Adaptation Scheme. *Proceedings ofNOSSDAV1998*.

Dr.Y.Madhavee Latha, presently working as a Principal at MRECW, Secunderabad. She received Ph.D (Signal Processing) from JNTU Hyderabad. She has more than 12 years of experience in the field of teaching. She contributed various research papers in journals & Conferences of national/international repute. Her area of interest includes multimedia systems, communication & signal processing. She is senior member of ISTE, IETE & IEEE, Technical societies.

Dr.D.Sreenivas Rao is professor in ECE department JNTU Hyderabad. He published many papers in international & national journals. He is a fellow of IETE and life member of ISTE Technical Society. His area of research in Advance Communication, Computer net works & network security.

Mrs. Vandana Khare is pursuing PhD in Communication Engineering JNTU Hyderabad (A.P). She completed M.E (Digital techniques) in 1999 from SGSITS, INDORE (M.P) India and B.E in ECE in the year 1994 from GEC Rewa (M.P). She is working as Associate Professor& Head ECE in Rishi Womens Engineering College, Hyderabad since July 2013. She has 17 years of teaching experience. She has published 08 research papers in International journals &presented o4 papers in National & International conferences. She is member of IEEE & life member of ISTE & IETE Technical societies. Her research Interest includes computer networks, mobile computing and Bio-medical imaging.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IMPLEMENTATION OF RADIAL BASIS FUNCTION AND FUNCTIONAL BACK PROPAGATION NEURAL NETWORKS FOR ESTIMATING INFRAME FILL STABILITY

[1]P.Karthikeyan and [2]S.Purushothaman

[1]P.Karthikeyan,
Research Scholar, Department of Civil Engineering,
CMJ University.

[2]Dr.S.Purushothaman,
Professor, PET Engineering College,
Vallioor-627117.

Abstract-**ANSYS 14 software is used for analyzing the infill frames. The numerical values of strain are used to train the artificial neural network (ANN) topology by using Back propagation algorithm (BPA) and Radial basis function network (RBF). The training patterns used for the ANN algorithms are chosen from the strain data generated using ANSYS program. During the training process, node numbers are presented in the input layer of the ANN and correspondingly, strain values are presented in the output layer of the ANN. Depending upon the type of values present in the patterns, the learning capability of the ANN algorithms varies.**

Keywords: Artificial Neural Network (ANN); Back propagation algorithm (BPA); Radial basis function (RBF).

I. INTRODUCTION

Structural material for building construction is based on masonry. Behavior of masonry under lateral load helps in evaluating the seismic vulnerability of existing buildings. Hence, proper retrofitting measures can be done. Masonry is held by the confining action of surrounding frame.The local behavior to masonry has to be considered during simulation. Masonry infills are a popular form of construction of high-rise buildings with reinforced concrete frames. The infilled frame consists of a moment resisting plane frame and infill walls. The masonry can be of brick, concrete units, or stones. Usually the RC frame is filled with bricks as non-structural wall for partition of rooms. Parking floor are designed as framed structures without regard to structural action of masonry infill walls. They are considered as non-structural elements. RC frames acts as moment resisting frames leading to variation in expected structural response. Masonry infill panels are treated as nonstructural element and their strength and stiffness contributions are neglected. The presence of infill wall changes the behavior of frame action into truss action thus changing the lateral load transfer mechanism.

II. RELATED WORK

Alirezaet al, 2013, gives a detailed presentation of a generic three-dimensional discrete-finite-element model that has been constructed for reinforced-concrete frames with masonry infill using ANSYS. Appropriate experimental data available from the literature are utilized to verify the model. The reasons behind some of previously observed damage to infill-frames are given. A simple method is proposed to overcome convergence issues which are related to the Newton–Raphson algorithm. It is shown that the model can be employed to predict the behavior of the infill-frame over a wide range of drift, and to interpret its response at various stages of in-plane or out-of-plane loading.

Cavaleriet al, 2013 treated the prediction of the response of infilled frames through the simplified approach of substituting the infill with an equivalent pin-jointed strut. In this framework, the results of an experimental study for the mechanical characterization of different types of masonry infills having the aim of estimating strength, Young modulus and Poisson's ratio are presented. Four types of masonry were investigated and subjected to ordinary compressive tests orthogonally to the mortar

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

beds and along the directions of the mortar beds. The experimental campaign confirmed the possibility of using an orthotropic plate model for prediction of the Poisson's ratio and Young modulus along the diagonal direction of infills (these parameters are requested by a model already known in the literature for the identification of struts equivalent to masonry infills). The experimental campaign made it possible to recognize a correlation between the Poisson's ratios and the strengths of masonries investigated along the orthotropic axes and to obtain the diagonal Poisson's ratio without specific experimental tests. Finally, the experimental responses of some infilled frames were used to test the reliability of the model proposed here.

Daniel et al, 2012 investigated the sensitivity of the seismic response parameters to the uncertain modelling variables of the infills and frame of four infilled reinforced concrete frames using a simplified nonlinear method for the seismic performance assessment of such buildings. This method involves pushover analysis of the structural model and inelastic spectra that are appropriate for infilled reinforced concrete frames. Structural response was simulated by using nonlinear structural models that employ one-component lumped plasticity elements for the beams and columns, and compressive diagonal struts to represent the masonry infills. The results indicated that uncertainty in the characteristics of the masonry infills has the greatest impact on the response parameters corresponding to the limit states of damage limitation and significant damage, whereas the structural response at the near-collapse limit state is most sensitive to the ultimate rotation of the columns or to the cracking strength of the masonry infills. Based on the adopted methodology for the seismic performance assessment of infilled reinforced concrete frames, it is also shown, that masonry infills with reduced strength may have a beneficial effect on the near-collapse capacity, expressed in terms of the peak ground acceleration.

III. PROBLEM DEFINITION

This paper provides neural network to supplement finite element analysis of infill masonry. During analysis strength and stiffness are evaluated subjected to lateral forces.

IV. MATERIALS AND METHODOLOGIES

A. Back Propagation Algorithm (BPA)

The concept of steepest-descent method is used in BPA [Atiya A.F., and Parlos A.G., 2000] to reach a global minimum. The number of layers are decided initially. The number of nodes in the hidden layers are decided. It uses all the 3 layers (input, hidden and output). Input layer uses 17 nodes, hidden layer has 2 nodes and the output layer includes two nodes. Flow-chart for BPA is shown in Figure 1.

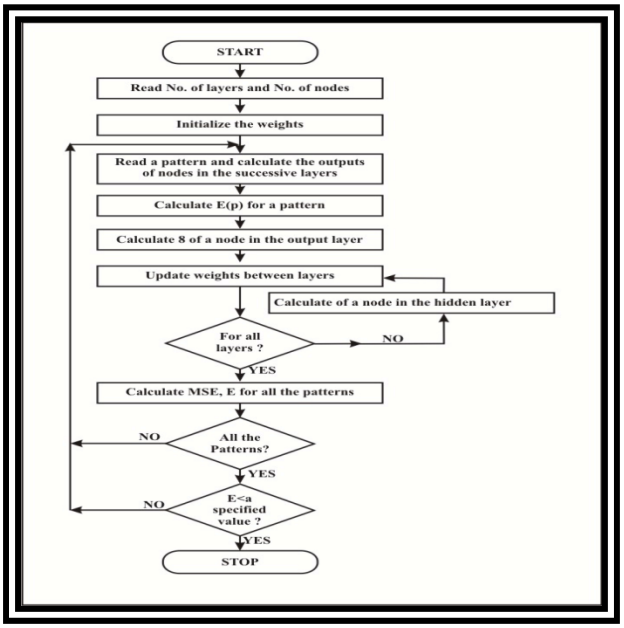


Fig.1 Flow-chart of BPA

STEPS INVOLVED IN

FORWARD PROPAGATION

The hidden layer connections of the network are initialized with weights.

The inputs and outputs of a pattern are presented to the network.

The output of each node in the successive layers is calculated.

$$O_{(\text{output of a node})}=1/(1+\exp(-\sum w_{ij}x_i)) \quad (1)$$

For each pattern error is calculated as:

$$E(p) = (1/2)\sum(d(p) - o(p))^2 \quad (2)$$

REVERSE PROPAGATION

For the nodes the error in the output layer is calculated

$$\delta_{(\text{outputlayer})}=o(1-o)(d-o) \quad (3)$$

The weights between output layer and hidden layer are updated

$$W_{(n+1)} = W_{(n)} + \eta\delta_{(\text{output layer})}\, o_{(\text{hidden layer})} \quad (4)$$

The error for the nodes in the hidden layer is calculated

$$\delta_{(\text{hidden layer})}= o(1-o)\sum\delta_{(\text{output layer})}W_{(\text{updated weights between hidden \& output layer})} \quad (5)$$

The weights between hidden and input layer are updated.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$W_{(n+1)} = W_{(n)} + \eta\delta_{(\text{hidden layer})}\, o_{(\text{input layer})} \quad (6)$$

The above steps complete one weight updation.

The above steps are followed for the second pattern for subsequent weight updation. When all the training patterns are presented, a cycle of iteration or epoch is completed. The errors of all the training patterns are calculated and displayed on the monitor as the MSE.

$$E_{(MSE)} = \sum E_{(p)} \quad (7)$$

B Radial Basis Function (RBF)

Radial Basis Functions (RBFs), [John Moody and Christian J Darken, 1989, Craddock R.J., and Warwick K., 1996], assures approximation to a function 'f' given input data. Input point is presented through a set of basis functions and the approximation is produced with RBF centers. Then the result is multiplied with a coefficient to sum them linearly.

The approximation is stored in the coefficients given function 't' to obtain centers of the RBF. RBFs have the following mathematical representation:

$$F(x) = c_0 + \sum_{i=0}^{N-1} c_i \Phi_a\left(\|x - CR_i\|\right) \quad (8)$$

Where,

c is a vector containing the coefficients of the RBF,

CR is a vector containing the center of the RBF, and

ϕ_a is the basis function or activation function of the network.

F(x) is the approximation produced as the output of the network. The coefficient c_o, which is a bias term, may take the value 0, if no bias is present. The norm used is the Euclidean distance norm. Equation (9) shows the Euclidean distance for a vector 'x' containing n elements:

$$\| x \| = \sqrt{\sum_{i=1}^{n} x_i^2} \quad (9)$$

Each center CR_j has the same dimension as the input vector 'x', which contains 'n' input values. The centers are points within the input data space and are chosen so that they are representative of the input data. When a RBF calculates its approximation to some input data point, the distance between the input point and each center is calculated, in terms of the Euclidean distance. The distances are then passed through the basis function ϕ_a. The results of the basis functions are weighted with the coefficients c_i and these weighted results are then linearly summed to produce the overall RBF output. One of the most common choices for the basis function is that of the Gaussian:

$$\Phi_a(x) = \exp\left(\frac{-x^2}{2\sigma}\right) \quad (10)$$

In which 'σ' is simply a scaling parameter. Other choices for the basis functions include the thin plate spline, the multi-quadric and the inverse multi-quadric.

Algorithm for Training RBF Network

There are two types of training techniques for RBF to choose the center values. They are:

1. Production of the coefficients using linear least squares optimization, after the center values have been chosen, using either some fixed or self organizing technique.

2. Production of the centers and coefficients using gradient descent equations. The center values are chosen during the training technique, instead of being fixed previously.

The following steps are the algorithm of training a radial basis function for the identification of electrical disturbances.

Step 1: Initialize the target values for the RBF network are used.

Step 2: Load the feature values

Step 3: Determine the RBF as,

$$K = X_{feat} \quad (11)$$

$$L = X_{feat} \quad (12)$$

$$\text{RBF} = \begin{bmatrix} e^{\left(-\sum_{j=1}^{n_{fe}}(K_{1j}-L_{1j})^2\right)} & \cdots & e^{\left(-\sum_{j=1}^{n_{fe}}(K_{1j}-L_{n_sj})^2\right)} \\ \vdots & \ddots & \vdots \\ e^{\left(-\sum_{j=1}^{n_{fe}}(K_{n_sj}-L_{1j})^2\right)} & \cdots & e^{\left(-\sum_{j=1}^{n_{fe}}(K_{n_sj}-L_{n_sj})^2\right)} \end{bmatrix} \quad (13)$$

Size of the RBF matrix is n_s x n_s.

Step 4: Calculate the matrix G as

$$G = \text{RBF} \quad (14)$$

$$A = G^T * G \quad (15)$$

Step 5: Compute the determinant value of matrix A. If it is zero do the step 6, otherwise go to step 7.

$$\text{Det} = |A| \quad (16)$$

Step 6: Find SVD of A. This process produces diagonal matrix S, and a decreasing order non-negative diagonal elements, and U and V unitary matrices, so that,

$$A = U * S * V^T \quad (17)$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Step 7: Compute the Inverse of matrix A

$$B = A^{-1} \quad (18)$$

Step 8: Determine the matrix E as

$$E = B * G^{T} \quad (19)$$

Step 9: Calculate the final weights and store in to file.

$$w_{RBF} = E * T_{RBF} \quad (20)$$

Step 10: Do the steps from 1 to 9 for remaining disturbance signals.

The flow chart for RBF implementation is given in Figure 2.

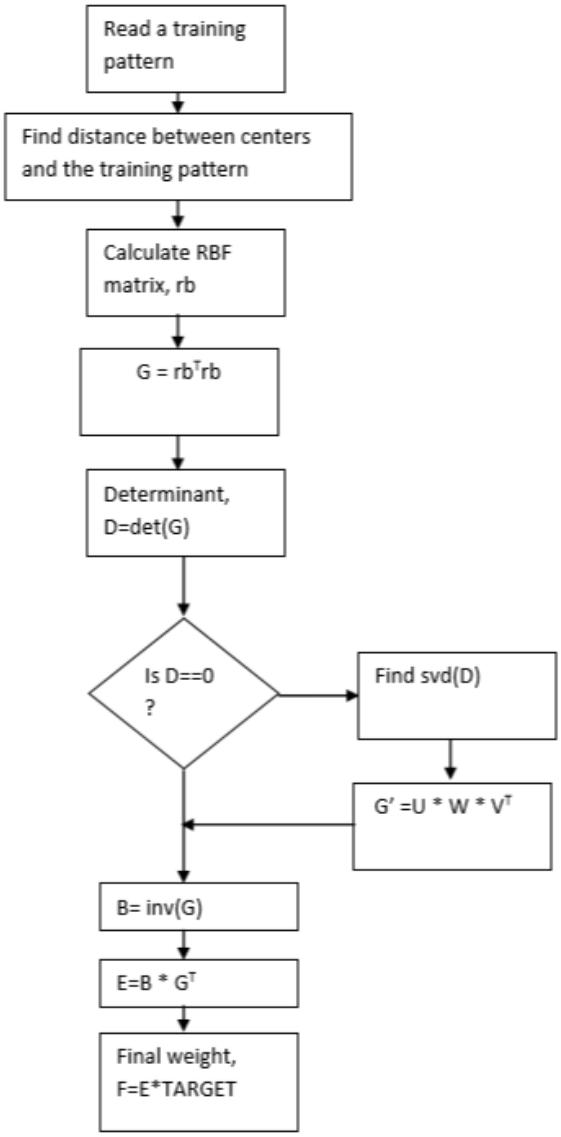


Fig. 2 Radial Basis Function Flow chart

V. EXPERIMENTAL SIMULATION

The model with struts fixed inside the frame is represented in Professional Engineer and imported as IGES into ANSYS 14

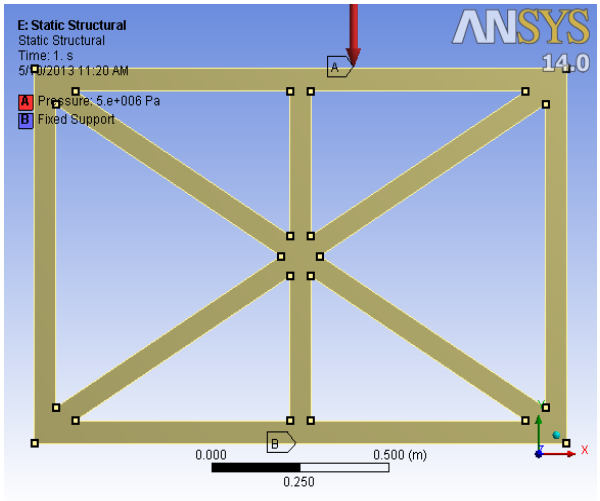


Fig.3 Model undeformed frame with struts as infill

TABLE.1 STRAIN-LIFE PARAMETERS

Strength Coefficient Pa	*Strength Exponent*	*Ductility Coefficient*	*Ductility Exponent*	*Cyclic Strength Coefficient Pa*	*Cyclic Strain Hardening Exponent*
9.2e+008	-0.106	0.213	-0.47	1.e+009	0.2

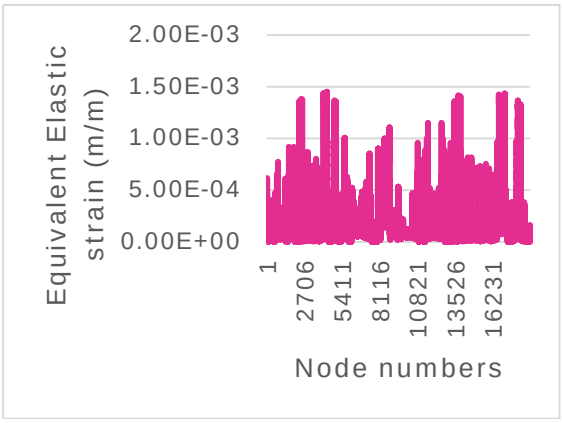


Fig.4 strain distribution

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fig. 4 shows the amount of strain presented at various nodes mentioned in the x-axis.

VI. RESULTS AND DISCUSSIONS

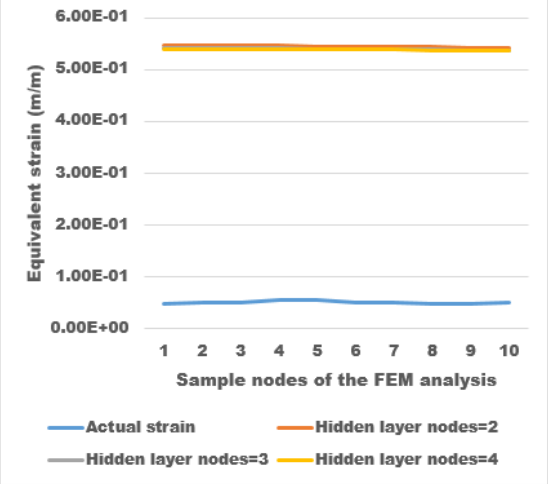


Fig.5 Estimation of strain by BPA

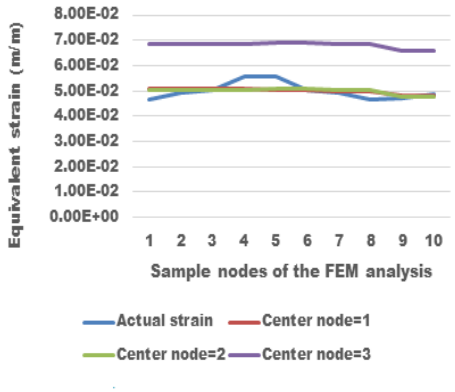


Fig. 6 Strain estimaion by RBF for different centers

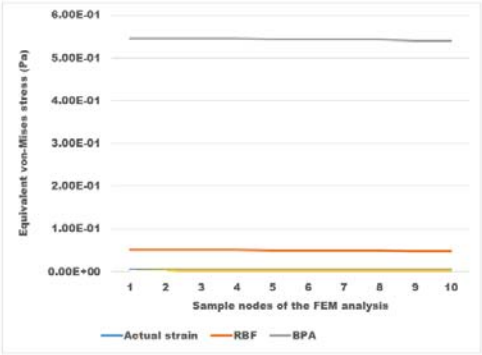


Fig.7 Performance comparison for estimation of strain data

The results has presented a detailed analysis of the performance of the proposed ANN algorithms for estimating the strength of infilled frame struts. The effect of numer of nodes in estimating the strain for the infilled frame is presented.

The algorithms are trained using 17 values in the input layer of the ANN topology and two values in the output layer: strain that is to be estimated during the testing stage of ANN algorithms. The number of nodes in the hidden layer for each algorithm varies depending upon the weight updating equations.

- As the number of training patterns increase, the time taken by BPA to converge also increases.
- The optimum number of nodes in the hidden used for BPA is 2 nodes. As the number of nodes in the hidden layer increases, the accuracy of BPA for stress estimation and strain estimation reduces.
- The optimum number of centers used for RBF in the hidden layer is 3. RBF does not require repeated learning of the strain patterns. In one iteration, the complete learning of the patterns presented in the input layer is achieved. RBF is best in estimating strain when compared to the performance evaluation of BPA.

VII. CONCLUSIONS

In this research work, ANSYS 14 software is used for analyzing the infill frames. A practical computer technique wherein each infil1 in a frame is replaced with an equivalent compression diagonal has been developed and tested for accuracy and reliability. Load deformation response characteristics of each infill in a frame are first determined using a finite element computer analysis. Results of this analysis are then used to develop the diagonal load deformation behavior of an equivalent diagonal brace which replaces the corresponding infill panel. In a

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

general analysis of a multi-panel structure, al1 panels are replaced by this method thus leading to a considerably simplified and economical technique.

REFERENCES

[1] AlirezaMohyeddina, Helen M. Goldsworthy, Emad F. Ga, 2013, FE modelling of RC frames with masonry infill panels under in-plane and out-of-plane loading, Engineering Structures, Vol.51, pp.73–87.

[2] Asteris, P.G., 2008, Finite Element Micro-Modeling of Infilled Frames, Electronic Journal of Structural Engineering, Vol.8, pp.1-11..

[3] Atiya A.F., and Parlos A.G., 2000, New results on recurrent network training: Unifying the algorithms and accelerating convergence, IEEE Trans. Neural Networks, Vol. \11, Issue 3, pp.697-709.

[4] Cavaleri L., Papia M., Macaluso G., Trapani F. Di, Colajanni P., 2013, Definition of diagonal Poisson's ratio and elastic modulus for infill masonry walls, Materials and Structures.

[5] Craddock R.J., and Warwick K., 1996, Multi-Layer Radial Basis Function Networks. An Extension to the Radial Basis Function, IEEE International Conference on Neural Networks, Vol.2, No.1, pp.700-705.

[6] Daniel Celarec, Paolo Ricci, MatjažDolšek, 2012, The sensitivity of seismic response parameters to the uncertain modelling variables of masonry-infilled reinforced concrete frames, Engineering Structures, Vol.35, pp.165–177.

[7] John Moody and Christian J Darken, 1989, Fast Learning in Networks of Locally-Tuned Processing Units, Neural Computation, Vol.1, No.2, pp.281-294.

P Karthikeyan is pursing PhD from CMJ University Meghalaya, India. He has 21 years of teaching experience in the Department of Civil Engineering. Presently he is working as Professor in SKP Engineering College. India

Dr.S.Purushothaman completed his PhD from Indian Institute of Technology Madras, India in 1995. He has 129 publications to his credit. He has 19 years of teaching experience. Presently he is working as Professor in PET Engineering College, India

Impact of Part-Of-Speech Based Feature Selection in Text Summarization

Rajesh Wadhvani
Computer Science Department
National Institute of Technology
Bhopal, India
Email: wadhvani_rajesh@rediffmail.com

R. K. Pateriya
Computer Science Department
National Institute of Technology
Bhopal, India
Email: pateriyark@gmail.com

Devshri Roy
Computer Science Department
National Institute of Technology
Bhopal, India
Email: devshriroy@manit.ac.in

***Abstract*—The standard practice in the field of summarization is to have a standard reference summary based on the queries. The summaries are manually generated by human experts. The automated summaries are then compared with the human generated summaries. In this paper a model which is useful for query-focused multi-document summarization is proposed. This model summarises documents of tagged data using cosine relevance measurement mechanism. Tagging is achieved by using the Stanford POS Tagger Package. We have used the concept of rarity in addition to traditional raw term frequency for assigning the weights to the features. In this work for a given sentence out of all possible tag sequences best is derived by using argmax computation. For evaluation DUC-2007 dataset is used. The summaries generated by our technique are compared with the stranded summaries provided by DUC 2007 using ROUGE (Recall-Oriented Understudy for Gisting Evaluation). The performance metrics used for compassion are Recall, Precision, F-score.**

Keywords: Text summarization, query-based summaries, sentence extraction.

I. INTRODUCTION

Summarizing the documents manually requires lots of efforts and time and thus it is very difficult. A technique is required where a computer program creates a shortened version of a text while preserving the content of the source, i.e., summarize document automatically. Automatic text summarization is not new. Work in this area originated in the 1950s, when the first breakthroughs were achieved by Luhn (1958). Despite this, most of the significant research in this area has been carried out in the last few years. The goal of automatic summarization is to take an information source, extract content from it and provide the most important content to the user in a condensed form, in a manner sensible to user's or application's need, see (Mani, 2001) for details. According to the function, a summary can be classified as generic or user-oriented. Generic summary present the authors viewpoint on the document. It considers all the information in the document to create summary. On the other hand user-oriented summary considers only that information which is relevant to user query. Based on relationship that a summary has to the source document, a summary can be either abstract or extract [1,2]. An extract involve selection of sentences or paragraphs from source document in the summary. On the other hand abstract involve the identification of salient concepts in the source document and rewrite them through natural language generation.

Words are classified into categories called part-of-speech. These are sometimes called word classes or lexical categories. These lexical categories are usually defined by their syntactic and morphological behaviours. The most common categories are nouns and verbs. Other lexical categories include adjectives, adverbs, prepositions and conjunctions. We have seen that words may belong to more than one part-of-speech (POS) category. A dictionary simple lists all possible grammatical categories for a given word. It does not tell us which word is used in which grammatical category in a given context. Grammatical categories, as indicated by POS tags, are rough indicators of meaning. For example "I banked the check, Thus bank if used as a noun indicates a financial establishment that invests money deposited by customers whereas the same word used as a verb may indicate activity to deposit money or valuables in a bank. Shop as a noun refers to the store where we go and buy things and same word when used as a verb refers to the activity of going to a shop, buying things, paying money etc. Thus knowing the grammatical category of word in context is helpful in ultimately determining the meaning. In this paper we uses POS based term to know that in general how a term is relevant and appropriate when it occurs in a language. This relevancy is checked with respect to POS Context.

In the process of summary generation the first major constituent is corpus analysis. Brute force linear scan approach of corpus analysis requires its own independent linear scan of all documents in the corpus. Moreover for every query it needs to repeat this linear scan of documents and would take huge amount of time. Clearly it is impractical. Better solution is somehow pre process the entire corpus in advance before you see any query and a priory organize the information about the occurrence of different words in the corpus in such a way that query processing can be much faster. When query is presented by user for summarization of documents, it is possible that query is not matched directly against the sentences of document. Only the parts of the query would be presented in document and sentences are retrieved. This is called problem of data spar city. So query processing is concerned with breaking the query into important parts, remove the suffixes

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

and obtain the root words. In the next step index table is constructed, terms which are meaningful called keywords or features comes under the index table. Index table consist of term frequency which is used to quantification of intra-document contents (similarity). For quantification of inter-document separation (dissimilarity) we use inverse-document frequency which is also the measure of rarity of the term in the collection. In the end, based on model sentence extraction from source documents is performed. In case of Boolean model, a set of sentences satisfying a Boolean query expression extracted from the document and the relevance of sentence with respect to query depends on how precisely Boolean query is expressed. It works on set theory which doesn't consider term frequency because for multiple occurrences of words, set theory only consider one term within set. In ranked retrieval model, query is free text query. The user query is one or more words in a human language where system returns an ordering over the (top) sentences in the collection.

This paper is organized as follows. Literature review is presented in Section 2. This chapter gives an account of previous works done in the area of text summarization. Chapter 3 and 4 discusses Part-of-Speech (POS) Tagging and Structure for representing the document respectively in detail. Description of the proposed work is given in Section 5. This chapter provides the detail of system architecture. Section 6 is Evaluation Procedure and Results, provides the description of tools used for evaluation and also gives the detail of evaluation metrics used for evaluation purpose. This section also provides the results obtained by the proposed methods. The experimental result shows the accuracy of the methods using precision, recall and F-score measures. Section 7 is the conclusion.

II. RELATED WORK

Earliest research instances on summarization proposed paradigms for extracting salient sentences from text using features like word and phrase frequency, position on the text and key phrases[1,2]. Understanding of contents, reformulation and sentence-compression is done while abstraction [4,5] where as the sentences of text are ranked and important ones are picked-up in extraction[3,4]. Luhn states that the frequency of a particular word in a given document provides degree of its worth i.e. how much worthy the word is in the document. As a first step document goes under the morphology analysis i.e. words are stemmed to their root forms, and stop words are deleted. Luhn then maintain a list of content words sorted by decreasing frequency, the index providing a significance measure of word.On a sentence level, a significance factor was derived that echo the number of occurrences of significant words within sentences, and the linear distance between them because of intervention of non-significant words. All sentences are ranked in order of their significance factor, and the top ranking sentences are finally selected to form the auto-summary.

Traditional information retrieval (IR) and its application like text summarization and text categorization techniques uses extensional vectorial representation [5,6], in which the original text is represented in the form of a numerical matrix. Matrix columns correspond to text sentences, and each sentence is represented in the form of a vector in the term space. These sentences may be compared based on various similarity measurement techniques. Christina Lioma[7] proposes a new type of term weight that is computed from part of speech (POS) n-gram statistics. The proposed POS-based term weight represents how informative a term is in general, based on the POS contexts in which it generally occurs in language. Latent Semantic Analysis (LSA) [8] and Singular Value Decomposition (SVD) [9] are used to find the generic summary of the text. LSA or SVD is applied to the matrix obtained to construct sentences representation in the topic space. The dimensionality of the topic space is much less than the dimensionality of the initial term space. The choice of the most important sentences is carried out on the basis of sentences representation in the topic space.

III. PART OF SPEECH (POS) TAGGING

Part of Speech Based Feature may improve the quality of summary. Various techniques for part of speech based term tagging based on Hidden Markov Model is discussed in[10,11,12]. Part of speech is a process that attaches each word in a sentence with a suitable tag from a given set of tags. The set of tags is called Tag-Set. In "People Jump high" for each word we have all possible tags(In our example three tags in Tag-Set). Many problems in AI fall into the class that is predict "Hidden" from "observed". In our case observations are words of given sentence and states are Tags.For each word we have a level with all possible tags and we want to find out best possible tag sequences T^*. Therefore whole graph is to be traversed for the best possible path from "" to "." which is based on argmax computation.Now

t_i = A particular Tag
T = Tag sequence
w_i = A particular word
W = word sequence

W: $w_1\ w_2\ w_3$ .....w_n
T: $t_1\ t_2\ t_3$ .....t_n

For given word sequence W and Tag sequence T if we place Tags corresponding to given word sequence, T* is best possible Tag sequence and our goal is maximize $P(T* \mid W)$ by choosing best T. Out of all possible Tag sequences best is derived by argmax computation. When we apply argmax on all possible Tag sequences with given word sequence it gives best possible Tag sequence T*. So best possible Tag sequence

T* = argmax $P(T \mid W)$
Where $P(T \mid W) = P(T_{1..n}|W_{1..n})$

$P(T \mid W) = P(T_1|W)\ .\ P(T_2|T_1W)\ .\ P(T_3|T_2T_1W)$ .....$P(T_n|T_{1..n-1}W)$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

By markov assumption for order one process a state only depends on previous state, hence

$P(T \mid W) = P(T_1|W) \ . \ P(T_2|T_1W) \ . \ P(T_3|T_2W)$ .....$P(T_n|T_{n-1}W)$

But these values cannot be computed easily because they are not available and it is difficult to find out these values.

argmax $P(T \mid W)$ = argmax {P(T) $P(W \mid T)$ / P(W)} Baye's theorem

Here P(W) is ignored because it is independent of T. Hence along with morkov assumption, baye's theorem is invoked which is very power full tool for problem solving in statistical AI(machine learning, NLP, Planning etc.).

argmax $P(T \mid W)$ = argmax {P(T) $P(W \mid T)$}

Now we introduce the Tags t_o and t_{n+1} as initial and final Tags respectively. Initial Tag is t_0 with probability one i.e, $P(t_0)$=1 and after tn the next Tag is t_{n+1} with probability one i.e, $P(t_{n+1}|t_n)$ =1. And w_0 is ϵ transition.

W: $\epsilon \ w_1 \ w_2 \ w_3$ .....w_n
T: $t_0 \ t_1 \ t_2 \ t_3$ .....$t_n \ t_{n+1}$

In above equation P(T) is Prior Probability, It acts as a nice filter to eliminate bad possibilities. P(T) is representation for highly likely Tag sequence as learned from the corpora. It help us to isolate bad Tag sequences and give weight to good Tag sequences.

P(T) = $P(t_0 = \wedge t_1 t_2 .....t_n t_{n+1} = .)$

P(T)= $P(t_0) \ P(t_1|t_0) \ P(t_2|t_1t_0)$ ..... $P(t_n|t_{n-1}t_{n-2}....t_0)$ $P(t_{n+1}|t_nt_{n-1}...t_0)$

By markov assumption for order one process(Bigram probability) that is a Tag only depends on previous Tag

P(T)=$P(t_0) \ P(t_1|t_0) \ P(t_2|t_1)$ ..... $P(t_n|t_{n-1}) \ P(t_{n+1}|t_n)$

P(T)= $\Pi_{i=1}^{n+1} \ P(t_i|t_{i-1})$

And $P(W \mid T)$ is likelihood probability of the word sequence for given Tag sequence T.

$P(W \mid T) = P(w_0|t_0....t_{n+1}) \ P(w_1|w_0t_0....t_{n+1})$ $P(w_2|w_1w_0t_0....t_{n+1})$.....$P(w_n|w_0...w_{n-1}t_0....t_{n-1})$

Lexical probability assumption that is word selection depends only on Tag chosen, hence

$P(W \mid T) = P(w_0|t_0) \ P(w_1|t_1) \ P(w_2|t_2)$.....$P(w_n|t_n)$ $P(w_{n+1}|t_{n+1})$

$P(W \mid T) = \Pi_{i=0}^{n+1} \ P(w_i \mid t_i)$

Now T^*= $[P(w_0|t_0) \ P(t_1|t_0)]$ $[P(w_1|t_1) \ P(t_2|t_1)]$ ..... $[P(w_n|t_n) \ P(t_{n+1}|t_n)]$ $[P(w_{n+1}|t_{n+1})]$

$T^* = \Pi_{i=0}^{n+1} \ P(t_i|t_{i-1}) \ P(w_i \mid t_i)$

where $P(t_i|t_{i-1})$ is Bigram probability
and $P(w_i|t_i)$ is Lexical probability
and $P(w_i|t_i)$=1 for i=0 {sentence beginner$\wedge$} and i=n+1 {full stop .}.

IV. DOCUMENT AND QUERY AS VECTOR

In v dimensional space, documents and query are presented by vectors where terms are axis of this space. Our aim is to rank documents according to their proximity to the query in this space. Here proximity means similarity and proximity (Similarity) between the vectors is equivalence to inverse of distance. Here distance is angle between the vectors of query and document. Angle between two vectors $\overrightarrow{a}$ and $\overrightarrow{b}$ is given by:

$\overrightarrow{a}.\overrightarrow{b} = |a||b|cos\theta$

Now take a document d and append it to itself and call this document $d^{'}$. Semantically d and $d^{'}$ have the same content but in weighted matrix each term of $d^{'}$ has double weight as compared to d.In vector space angle between document d and $d^{'}$ is zero. And when we want to compare this documents d and $d^{'}$ with query q: angle between q and d and also between q and $d^{'}$ is same. If two documents have same proportion of words whether they are longer or shorter, have identical vectors after length normalization. A vector can be length normalized by dividing each of its components by its length for this we use L2 norm:

$\| \overrightarrow{x} \|_2 = \sqrt{\sum {x_i}^2}$
$\| \overrightarrow{d} \|_2 = \sqrt{{x_1}^2 + {y_1}^2 + {z_1}^2}$
$\| \overrightarrow{d'} \|_2 = \sqrt{4{x_1}^2 + 4{y_1}^2 + 4{z_1}^2}$

Dividing a vector by its L2 norm makes it a unit vector. Now

$\frac{\overrightarrow{d}}{\| \overrightarrow{d} \|_2} = \frac{X_1}{\sqrt{{X_1}^2+{Y_1}^2+{Z_1}^2}}\hat{i} + \frac{y_1}{\sqrt{{X_1}^2+{Y_1}^2+{Z_1}^2}}\hat{j} + \frac{Z_1}{\sqrt{{X_1}^2+{Y_1}^2+{Z_1}^2}}\hat{k}$

$\frac{\overrightarrow{d'}}{\| \overrightarrow{d'} \|_2} = \frac{2X_1}{2\sqrt{{X_1}^2+{Y_1}^2+{Z_1}^2}}\hat{i} + \frac{y_1}{2\sqrt{{X_1}^2+{Y_1}^2+{Z_1}^2}}\hat{j} + \frac{Z_1}{2\sqrt{{X_1}^2+{Y_1}^2+{Z_1}^2}}\hat{k}$

$\frac{\overrightarrow{d}}{\| \overrightarrow{d} \|_2} = \frac{\overrightarrow{d'}}{\| \overrightarrow{d'} \|_2}$ normalized vectors

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$Length(\frac{\overrightarrow{d}}{\|\overrightarrow{d}\|_2}) = \frac{{X_1}^2}{{X_1}^2+{Y_1}^2+{Z_1}^2} + \frac{{Y_1}^2}{{X_1}^2+{Y_1}^2+{Z_1}^2} + \frac{{Z_1}^2}{{X_1}^2+{Y_1}^2+{Z_1}^2} = 1$ unit vector

After converting the document in its unit vector we can calculate the angle between the query and document. Actually we are interested in finding the distance between query vector and normalized document vector. Angle between two vectors $\overrightarrow{a}$ and $\overrightarrow{b}$ is given by

$$\overrightarrow{a}.\overrightarrow{b} = |a||b|cos\theta$$

if $\overrightarrow{a}$ and $\overrightarrow{b}$ are normalized then

$$\overrightarrow{a}.\overrightarrow{b} = cos\theta$$

Now angle between Query vectors $\overrightarrow{q}$ and Document vector $\overrightarrow{d}$ where $\overrightarrow{q}$ and $\overrightarrow{d}$ are normalized unit vectors, is given by

$$cos(\overrightarrow{q}.\overrightarrow{d}) = \overrightarrow{q}.\overrightarrow{d}$$

V. PROPOSED MODEL

We propose a model which is useful for query-focused multi-document summarization. This model summarizes documents using the Relevance Measurement Mechanism. The metric used for relevance is, tf*ipf, where tf stands for term frequency and ipf stands for inverse Para frequency. We include Inverse Paragraph frequency as a rarity measure of each term. This rarity is a metric to identify the importance of the term. It works based on the assumption that a word which is rarer, should have more importance, and hence more weight age in the score calculation. The proposed methods needs the proximity of documents and query, and provide a normalize weight to each document associated with a single query and later uses this weight to rank sentences of each document. Our method takes advantage of the fact that the more closely the document with query, the likelihood of its sentences close to query is higher. This method works on tagged data for better and more accurate summarization. Tagging is achieved by using the Stanford POS Tagger Package. A POS tagger is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc. The architecture and implementation of the proposed system is based on retrieving information from multiples document and form a summary. The complete operational process can be described in following phases where initially the multiple text documents and query is given as the input.

i) Document pre-processing
ii) Part-of-speech tagging
iii) Document presentation using paragraph segmentation
iv) Matrix formulation and normalization using tf-ipf weighting scheme
v) Sentence Scoring

(i) Document pre-processing

This phase is for both the query and documents. First of all, the topics file which contains topic or query for each set of documents is read. After fetching the query and the document, these documents are passed through a pre-processor where stop-words and special symbols like punctuation marks are removed to process them into standard text file format. After this, the pre processed files are passed through a tagger to tag all parts of speech. Such a tagged file keeps track of all words, as well as their forms of speech.

(ii) Sentence POS Tagging

This is speech identification phase. Part of speech tags give concerning information about the role of a word in its lexicon. It may also provide information about the verbalization of a word. POS tags are a valuable part of a language processing tasks; they provide favourable information to other components such as a parser or a named-entity recognizer. Most of the POS tagging is done by Brills transformation based learning which take unidirectional approach. Here for tagging unidirectional approach is used. New tag (t_0) is figured considering the previous tag (t_{-1}), for finding tag (t_1), tag (t_0) is considered; this is done in backward direction. Each sentence POS tagging is performed using Stanford POS tagger [13].

For sequences of word in sentence $w_1, w_2, w_3, ..., w_n$ a sequence of tags $t_1, t_2, t_3, ..., t_n$ are assigned respectively based on Peen Treebank tag dictionary.

Event Sub sentence is generated with word w_i and its corresponding tag t_i

$S[(w_1, t_1), (w_2, t_2)....(w_i, t_i)]$

POS Tagging Algorithm-

i) Initialize an object of MaxentTagger giving the corpus to be used as the argument.
ii) Read the file to be tagged.
iii) Create a new file called out.txt to store the tagged text.
iv) Split the text read into paragraphs. para=untagged.split();
v) Loop through all paragraphs. for i=1 to n do tagged=tagString(para[i]) out.append(tagged) end

(iii) Document presentation using paragraph segmentation

After the raw files are prepared, these are passed through a structuring phase, where all the files are structured into one unique term verses paragraph matrix. Each document Dm consist of different paragraphs named as $p_1, p_2, ..., p_n$ where each paragraph is collection of m unique terms $t_1, t_2, t_3, ...t_m$.

Q = $(t_1, t_2, t_3...t_m)$ is set of unique terms.

D = $(D_1, D_2, D_3...D_n)$ is set of documents associated with query.

S = $((p_{1,1}, p_{2,1}, ..., p_{j,1})), (p_{1,2}, p_{2,2}, ..., p_{k,2}), ..., (p_{1,n}, p_{2,n}, ..., p_{l,n}))$ is the set of sets, and each set is the paragraph collection per document that is, s_l, n is the lst paragraph of document n.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

(iv) Matrix formulation and normalization using tf-ipf weighting scheme

In vector model each document and query is represented by a weighted vector in dimensions of terms. Terms which are meaningful comes under the index table and form a term-document weighted matrix. Two types of weights are incorporated for each term which are:

i) Local term weighting (Log-frequency weighting): It measures the importance of a term within a document. A document or zone that mentions a query term more often has more to do with that query. Therefore weight of index term should consider term frequency. Number of occurrence of term within document is known as raw term frequency, but it is not what we want because relevance does not increase proportionally with raw term frequency. One well-studied technique is use log-frequency weighting. We can compute a log-frequency weight for each term t in document d by

$$w_{(t,d)} = \begin{cases} 1 + log_{10} tf_{(t,d)} & if tf_{(t,d)} > 0, \\ 0 & Otherwise \end{cases}$$

ii) Global term weighting (inverse sentence frequency): Raw term frequency as above suffers from a critical problem: all terms are considered equally important when it comes to assessing relevancy on a query. It does not consider the rarity of the term. Rarity of the term is also important because rare terms in a collection are more informative than frequent terms.
So we need a measurement for rarity of the term within the collection.Collection may be a collection of all documents or all paragraphs or all sentences of corpus. In field of Information retrieval where we have large number of document in corpus, Document frequency is inverse measure of the informativeness of term within the document collection. But in case of text summarization where numbers of documents are limited and are very less in number, inverse sentence frequency is better measurement to check the rarity of the terms. The sentence frequency sf_t, defined to be the number of documents in the collection that contain a term t. If N is total number of sentences in documents collection then $sf_t \leq N$. We define the inverse sentence frequency (isf) of t by

$$isf_t = log_{10}(N/sf_t)$$

We use $log_{10}(N/sf_t)$ instead of N/sf_t to "dampen" the effect of isf. Inverse sentence frequency proves that "Rare words are more informative as compared to frequent words".

(v) Sentence Scoring

After both the data and the query have been structured, they are passed through a similarity calculation phase, where dot product of each paragraph with the query is calculated. This product indicates the score of each paragraph. Finally, a given threshold of paragraphs is fetched according to their scores, arranged in descending order. This combination of paragraphs results in our final summary.

VI. EVALUATION PROCEDURE AND RESULTS

(i) Test Dataset: For evaluation DUC-2007 dataset is used and it is available through [14] on request. There are 45 topics in the dataset and for each topic a set of 25 relevant documents are given. Each DUC topic comprises of four part; document set number, title of topic, narration and the list of document associated with topic. In this paper, the narration part of topic is used to frame the query. Table 1 shows the description of DUC-2007 dataset.

TABLE I: Dataset description

Dataset description	DUC-2007 dataset
Number of topics	45
Number of collections	45
Number of documents per collections	25
Total number of documents in dataset	$45 * 25$
Summary Length	250 words

(ii) Evaluation Metrics: The standard practice in the field of summarization is to have a standard reference summary based on the queries. The summaries are manually generated by human experts. The automated summaries are then compared with the human generated summaries. Evaluation results are normally obtained by the ROUGE (Recall-Oriented Understudy for Gisting Evaluation). It is a summary evaluation package for judging the performance of the summarization system. The ROUGE summary evaluation package is written in Perl and it is available through [15].To evaluate the accuracy and relevance of the automated summary with respect to the expert summaries, three metrics are used :

i) Recall
ii) Precision
iii) F-score

F-measure is a measure of a system's summary accuracy. It considers both the precision p and the recall r of the system's summary to compute the score. Precision reflects how many of the system's extracted sentences are relevant, and Recall reflects how many relevant sentences the system missed.

Given an input text, a expert's summary, and a automated summary, these scores inform us by quantifying that how closely the system's summary corresponds to the human one. For each unit, we let correct = the number of sentences extracted by the system and the human; wrong = the number of sentences extracted by the system but not by the human; and missed = the number of sentences extracted by the human but not by the system. Then

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$Precision = correct/(correct + wrong)$$

$$Recall = correct/(correct + missed)$$

$$F - Score = \frac{(1+\beta^2)Recall*Precision}{Recall+\beta^2 Preicision}$$

Where, F_β "measures the effectiveness of system's summary with respect to a user who attaches β times as much importance to recall as precision".

(iii) Evaluation Procedure: On each document this method is performed and for tagging Stanford tagger is used. Using the scoring scheme we found that most important sentences are at the top most of the time. Here we rely on the results of part-ofspeech tagger. The standard practice in the field of summarization is to have a standard reference summary based on the queries. The summaries are manually generated by human experts. The automated summaries are then compared with the human generated summaries. Evaluation results are normally obtained by the ROUGE (Recall-Oriented Understudy for Gisting Evaluation). It is a summary evaluation package for judging the performance of the summarization system.
The automated summaries are compared with the available reference summaries and evaluation results are obtained by the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) summary evaluation package for judging the performance of the summarization system. To compute ROUGE-Scores, ROUGE-1.5.5 will be run with the following parameters :
ROUGE-1.5.5.pl -n 4 -2 -1 -U -c 95 -r 1000 -f A -p0.5 -t 0 -s settings.xml
Where, settings.xml is a xml file for specifying system summaries and corresponding reference summaries locations.

TABLE II: Results using Average model scoring formula for untagged corpus

Evaluation Method	Average_R	Average_P	Average_F
ROUGE-1	0.31558	0.38584	0.32977
ROUGE-2	0.07226	0.08086	0.06924
ROUGE-3	0.02354	0.02481	0.02105
ROUGE-4	0.01161	0.01188	0.00998
ROUGE-L	0.28854	0.35203	0.30069
ROUGE-W-1.2	0.08427	0.19030	0.10961
ROUGE-S*	0.10840	0.15389	0.10696
ROUGE-SU*	0.10676	0.15136	0.10504

TABLE III: Results using Average model scoring formula for tagged corpus

Evaluation Method	Average_R	Average_P	Average_F
ROUGE-1	0.30193	0.40529	0.34105
ROUGE-2	0.06460	0.08716	0.07321
ROUGE-3	0.02070	0.02857	0.02372
ROUGE-4	0.00982	0.01388	0.01137
ROUGE-L	0.27432	0.36959	0.31045
ROUGE-W-1.2	0.08040	0.20178	0.11362
ROUGE-S*	0.09191	0.16427	0.11207
ROUGE-SU*	0.09024	0.16158	0.11006

TABLE IV: Results using Best model scoring formula for untagged corpus

Evaluation Method	Average_R	Average_P	Average_F
ROUGE-1	0.34963	0.42061	0.36337
ROUGE-2	0.09591	0.11115	0.09452
ROUGE-3	0.04022	0.04576	0.03847
ROUGE-4	0.02406	0.02689	0.02257
ROUGE-L	0.32016	0.38687	0.33304
ROUGE-W-1.2	0.09450	0.21379	0.12298
ROUGE-S*	0.12546	0.17533	0.12404
ROUGE-SU*	0.12727	0.17793	0.12610

TABLE V: Results using Best model scoring formula for tagged corpus

Evaluation Method	Average_R	Average_P	Average_F
ROUGE-1	0.33409	0.43803	0.37278
ROUGE-2	0.08903	0.12100	0.10131
ROUGE-3	0.03761	0.05316	0.04351
ROUGE-4	0.02159	0.03119	0.02518
ROUGE-L	0.30401	0.40317	0.34097
ROUGE-W-1.2	0.09043	0.22214	0.12659
ROUGE-S*	0.10878	0.18927	0.13037
ROUGE-SU*	0.11060	0.19065	0.13215

VII. Conclusion

The paper described the model for extracting paragraphs from POS based tagged corpora using cosine similarity measurement mechanism. Initial investigation shows that applying cosine similarity on tagged or untagged corpora the results are more or less same with marginal difference. As in case of untagged corpora the result(F-Score) is 0.36337 and with tagged corpora result(F-Score) is 0.37278. In future we will focus on developing some other similarity based measurements that will support the POS based features for enhancing the performance of text summarization.

References

[1] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", In Journal of Emerging Technologies in Web Intelligence, pp. 258-268, Vol 2, No 3 (2010), Aug 2010.
[2] K. Knight and D.Marcu, "Summarization beyond sentence extraction: a probablistic approach to sentence compression", Artefcial Intelligence, pages 91-107, 2002 Elsevier Science.
[3] D. Zajic, B. J. Dorr, J. Lin, and R. Schwartz, "Multi-candidate reduction: Sentence compression as a tool for document summarization tasks" , Inf. Process. Manage, Volume 43, pp. 1549-1570, November 2007.
[4] H. Daume, D. Marcu, "A noisy-channel model for documentcompression", In proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Ser. ACL 02. Stroudsburg, PA, USA:Association for Computational Linguistics, pp. 449-456, 2002.
[5] Nicola Polettini, "The Vector Space Model in Information Retrieval- Term Weighting Problem", http://sra.itc.it/people/polettini/PAPERS/Polettini Information Retrieval.pdf, 2004.
[6] Johannes Furnkranz, "A Study Using n-gram Features for Text Categorization", Technical Report OEFAI-TR-98-30, Austrian Research Institute for Artificial Intelligence, Austria, 1998.
[7] Christina Lioma and Roi Blanco, Part of Speech Based Term Weighting for Information Retrieval, Lecture Notes in Computer Science Volume 5478, 2009, pp 412-423

[8] I. V. Mashechkin, M. I. Petrovskiy, D. S. Popov, and D. V. Tsarev, "Automatic Text Summarization Using Latent Semantic Analysis". In Journal of Programming and Computer Software, pp. 299305, 2011, Vol. 37, No. 6.
[9] Josef Steinberger and Karel Jezek, " Text Summarization and Singular Value Decomposition", Lecture Notes in Computer Science Volume 3261, 2005, pp 245-254.
[10] Zhang Youzhi, "Research and Implementation of Part-of-Speech Tagging based on Hidden Markov Model",Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications: IEEE, 2009,978-1-4244-4607-0/09.
[11] Sang-Zoo Lee and Jun-ichi Tsujii, Hae-Chang Rim.,"Lexicalized Hidden Markov Models for Part-of-Speech Tagging", ACM DIGITAL LIBRARY., Proceedings of the 18th conference on Computational linguistics - Volume 1, 2000, pp.481-487 .
[12] Scott M. Thede and Mary P. Harper.,"A second-order Hidden Markov Model for part-of-speech tagging", ACM DIGITAL LIBRARY., Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp.175-182.
[13] http://nlp.stanford.edu/software/tagger.shtml
[14] http://duc.nist.gov/data.html
[15] http://www.berouge.com/Pages/DownloadROUGE.aspx.

AUTHORS PROFILE

Prof. Rajesh Wadhvani B.E in Computer Science from Rajiv Gandh Technical University,M.Tech in Computer Science from Maulana Azad National Institute of Technology Bhopal, Persuing PhD in Computer science from Maulana Azad National Institute of Technology Bhopal. Presently Working as Asst. Prof in Department of Information Technology in Maulana Azad National InstituteTechnology, Bhopal.

Dr. R. K. Pateriya Ph.D from Maulana Azad National Institute Technology, Bhopa. Presently Working as Associate Prof. in Department of Information Technology in Maulana Azad National Institute of Technology, Bhopal.

Dr. Devshri Roy Ph.D from IIT Kharagpur, Specialization in Application of Computer and Communication Technologies in E-learning , Personalized Information Retrieval , and Natural Language Processing. Presently Working as Associate Prof. in Department of Information Technology in Maulana Azad National Institute of Technology, Bhopal.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

DETECTION OF MICROCALCIFICATION IN BREAST USING RADIAL BASIS FUNCTION

[1]Shaji B., [2]Purushothaman S., and [3]Rajeswari R.,

[1]Shaji B.,
Research Scholar,
Vels University, Pallavaram, Chennai, India-600117.

[2]Dr.Purushothaman S.,
Professor,
PET Engineering College, Vallioor, INDIA-627117,

[3]Rajeswari R.,
Research scholar,
Mother Teresa Women's University, Kodaikanal-624102, INDIA.

***Abstract*-This paper presents combination of wavelet with radial basis function (RBF) in identifying the microcalcification (MC) in a mammogram image. Mammogram image is decomposed using Coiflet wavelet to 5 levels. Statistical features are extracted from the wavelet coefficients. These features are used as inputs to the RBF neural network along with a labeling of presence or absence of MC. The classification performance of RBF is minimum 95% out of the presence of total MC in a given mammogram**.

Keywords: mammogram , Microcalcification, Radial basis function, Coiflet wavelet

I. INTRODUCTION

Ultrasound (US) is a useful diagnostic tool to distinguish benign [Veldkamp, et al, 2000] from malignant[Ackerman et al, 1973, Jiang et al, 1996] masses of the breast. It is a very convenient and safe diagnostic method. However, there is a considerable overlap of benignancy and malignancy in ultrasonic images and interpretation is subjective. Mammogram imaging [Bovis,et al,2000, Leichter et al, 2000] is ideal and indispensable for women older than 40 years, for whom the risk of breast cancer is increased. Most breast disorders are not cancer, and even in the remaining number of cancer cases, more than 90% are curable, if detected early and promptly treated. Mammograms[Dhawan et al, 1996, Cheng et al, 2003, Sampat, et al, 2005] are not 100% accurate, scheduling a regular mammogram represents the best radiological way to find breast changes early before there are any obvious signs or symptoms of cancer.

II. LITERATURE SURVEY

Dehghan and Abrishami, 2008 present a computer-aided diagnosis (CAD) system for automatic detection of clustered MCs in digitized mammograms. The method is applied to a database of 40 mammograms (Nijmegen database) containing 105 clusters of MCs.

Madabhushi and Metaxas, 2003, present a technique to automatically find lesion margins in ultrasound images, by combining intensity and texture with empirical domain specific knowledge along with directional gradient and a deformable shape-based model. The images are first filtered to remove speckle noise and then contrast enhanced to emphasize the tumor regions. Probabilistic classification of image pixels based on intensity and texture is followed by region growing using the automatically determined seed point to obtain an initial segmentation of the lesion..

Chang et al., 2005, stated that the tumors are segmented using the newly developed level set method at first and then six morphologic features are used to distinguish the benign and malignant cases. The support vector machine (SVM) is used to classify the tumors. There are 210 ultrasonic images of pathologically proven benign breast tumors from

.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

120 patients and carcinomas from 90 patients in the ultrasonic image database.

O'Neill and Penm, 2007, have used subset polynomial neural network techniques in conjunction with fine needle aspiration cytology to undertake this difficult task of predicting breast cancer.

Liao, et al. 2011, established a set of features for differentiating benign from malignant breast lesions US images. Two types of features (sonographic and textural features) are considered. Among them, three sonographic features are novel. Sonograms of 321 pathologically proven breast cases are analyzed and classified into benign and malignant categories. The discrimination capability of the extracted features are evaluated using the support vector machines (SVM) in comparison with the results obtained from artificial neural networks (ANN) and K-nearest neighbor (KNN) classifier..

Chunekar and Ambulgekar, 2009, highlights on different neural network approaches to solve breast cancer problem. They emphasized on the use of Jordan Elman neural network approach on three different database of breast cancer.

Mahjabeen and Monika, 2012, studied various techniques used for the diagnosis of breast cancer. It was found that the combination of Artificial Neural Networks in most of the instances gives accurate results for the diagnosis of breast cancer and their use can also be extended to other diseases.

III. MATERIALS AND METHODOLOGY

A. MIAS DATABASE

Table 1 MIAS image details

Image number	Appearance	Condition	B/M	X, Y	Radius of pixels
Mdb045	Fatty-glandular	Normal			
Mdb046	Fatty-glandular	Normal			
Mdb047	Fatty-glandular	Normal			
Mdb049	Fatty-glandular	Normal			
Mdb050	Fatty-glandular	Normal			
Mdb051	Fatty-glandular	Normal			
Mdb053	Dense-glandular	Normal			
Mdb054	Dense-glandular	Normal			
Mdb055	Fatty-glandular	Normal			
Mdb169	Dense-glandular	Normal			
Mdb209	Fatty-glandular	Calcification	M	647, 503	87
Mdb210	Fatty-glandular	Normal			
Mdb211	Fatty-glandular	Calcification	M	680, 327	13
Mdb213	Fatty-glandular	Calcification	M	547, 520	45
Mdb214	Fatty-glandular	Calcification	B	582, 916	11
Mdb215	Dense-glandular	Normal			
Mdb216	Dense-glandular	Calcification	M	480, 794	60
Mdb219	Fatty-glandular	Calcification	B	546, 756	29
Mdb221	Dense-glandular	Normal			
Mdb222	Dense-glandular	Calcification	B	398, 427	17
Mdb223	Dense-glandular	Calcification	B	591, 529	6
Mdb224	Dense-glandular	Normal			
Mdb226	Dense-glandular	Calcification	B	287, 610	7
Mdb227	Fatty-glandular	Calcification	B	504, 467	9
Mdb233	Fatty-glandular	Calcification	M	630, 270	83
Mdb237	Fatty	Normal			
Mdb238	Fatty	Calcification	M	522, 553	17
Mdb239	Dense-glandular	Calcification	M	567, 808	25
Mdb240	Dense-glandular	Calcification	B	643, 614	23
Mdb241	Dense-glandular	Calcification	M	453, 678	38
Mdb242	Dense-glandular	Normal			
Mdb245	Fatty	Calcification	M	625, 197	89
Mdb246	Fatty	Normal			
Mdb247	Fatty	Normal			
Mdb248	Fatty	Calcification	B	378, 601	10
Mdb249	Dense-glandular	Calcification	M	575, 639	64
Mdb250	Dense-glandular	Normal			
Mdb252	Fatty	Calcification	B	439, 367	23
Mdb253	Dense-glandular	Calcification	M	733, 564	28
Mdb322	Dense-glandular	Normal			

.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

B.WAVELET FEATURES EXTRACTION

Wavelet [Wang and Karayiannis, 1998, Jamarani et al, 2005] decomposition is introduced as the main method for feature generation. The Wavelet [Strickland and Hahn, 1996] coefficients are divided into low (L) frequency approximation coefficients and high (H) frequency detail-coefficients. The high frequency coefficients are further divided into detail sub-bands: vertical (LH), horizontal (HL), and diagonal (HH) coefficients. The low frequency (LL) approximation-coefficients provide a reduced resolution representation of the original image which can be transformed again according to the wavelet level applied. Applying wavelet decomposition to an image will produce an approximation matrix that is a quarter of the original area of an image.

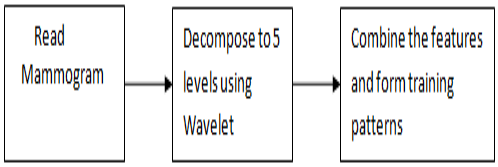


Fig.1 Feature extraction using Wavelet

The features are obtained from the Approximation and Details of the 5th level by using the following equations

V1=1/d ∑ (Approximation details)

Where d = Intensity values and

V1 = Mean value of approximation

V2=1/d ∑ (Approximation or details –V1))

Where V2=Standard Deviation of approximation

V3=maximum (Approximation or details)

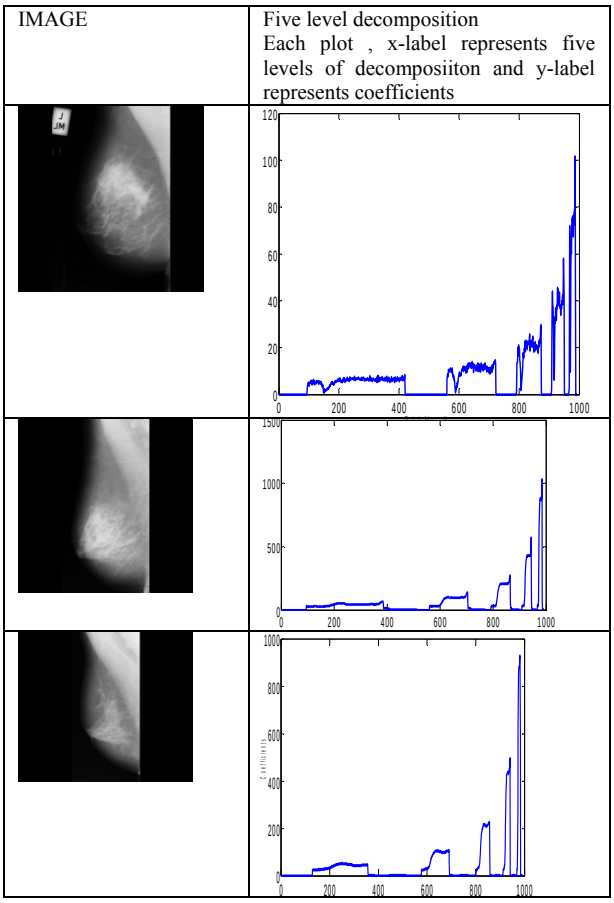
V4=minimum (Approximation or details)

V5=norm (Approximation or Details)^2

Where V5 = Energy value of frequency

The flow of Feature extraction using Wavelet is shown in Figure 1. Mammogram image is given as input to the system and level 1 to level 5 decompositions take place. Initially, Approximation, horizontal, vertical and diagonal matrices are obtained from the original image. Each matrix is ¼ th size of the input image. In the level two and subsequent levels, Approximation matrix of the previous levels are used for subsequent decompositions.

TABLE 2 MAMMOGRAM AND ITS FEATURES



C RADIAL BASIS FUNCTION (RBF)

An RBF [Tsujii, et al, 1999] neural network consists of an input and output layer of nodes and a single hidden layer. Each node in the hidden layer implements a basis function $G(\mathbf{x}\mathbf{x}_i)$ and the number of hidden nodes is equal to the number of data points in the training database. Radial basis function is a supervised neural network[Gholamali and Jamarani, 2005, Papadopoulos et al, 2005].

Training RBF is done as follows,

Step 1: Finding distance between training pattern and centers.

Step 2: Creating an RBF matrix whose size will be (np X cp), where np = number of roughness patterns (100 patterns X number of mammogram) used for training and cp is number of centers which is equal to 100. The number of centers chosen should make the RBF network learn the maximum number of training patterns under consideration.

Step 3: Calculate final weights which are inverse of RBF matrix multiplied with Target values.

.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Step 4: During testing the performance of the RBF network, RBF values are formed from the features obtained from the image and processed with the final weights obtained during training. Based on the result obtained, the image is classified to have MC or no MC.

Testing RBF

Step 1:.Read the features

Step 2: Read the final weights

Step 3 Calculate.

Numerals = F * E

Step 4: Check the output with the MC template.

IV. RESULTS AND DISCUSSIONS

Figure 1 shows original mammogram of a breast. Figures 2-6 show decomposed images. In each image, the approximation, horizontal, vertical and diagonal have been combined and displayed.

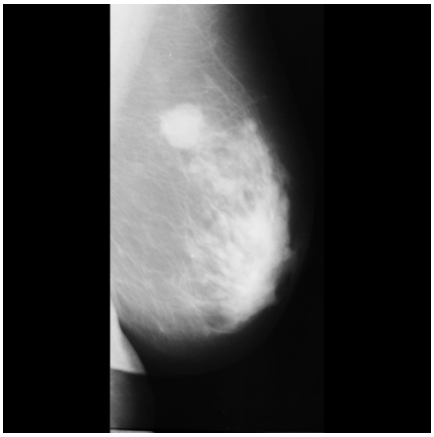


Fig. 1 Original mammogram image

Figure 2 shows the default view of wavelet decomposition with the features extracted from the top left corner, which is the low frequency image from the second level of decomposition. Figures 3-6 show all the images from each level separately and at the same size. This provides a better view of the differences between levels of decomposition especially when looking at the low frequency image that produces the approximation coefficients for generated features.

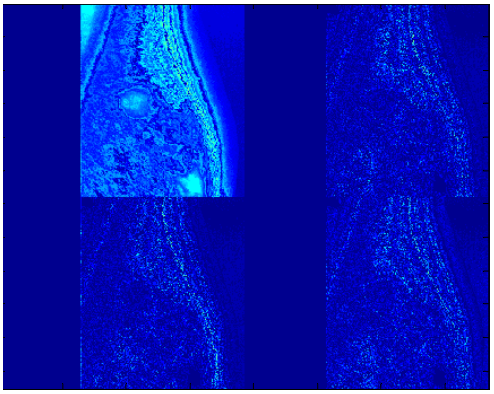


Fig. 2 Level-1 decomposition

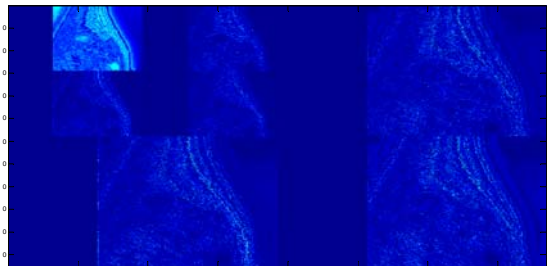


Fig. 3 Level-2 decomposition

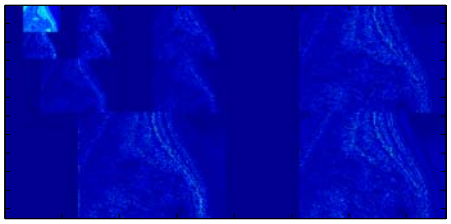


Fig. 4 Level-3 decomposition

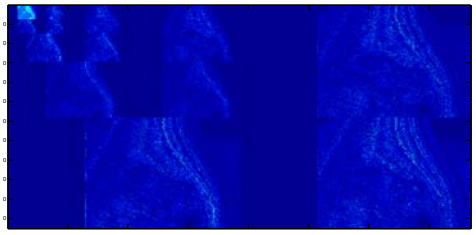


Fig. 5 Level-4 decomposition

.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

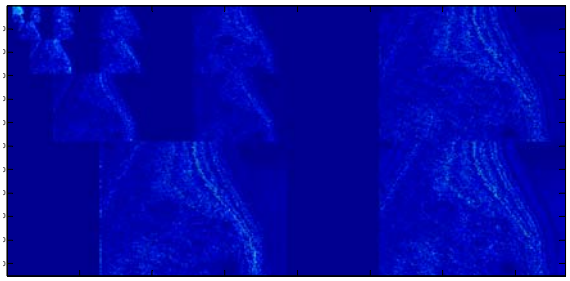


Fig.6 Level-5 decomposition

Figures 7-10 show coefficients of all levels for approximaiton, Horizontal, Vertical and Diagonal. These coefficients are used as inputs for equations 1-5 for generating patterns. These patterns are used for training the RBF network.

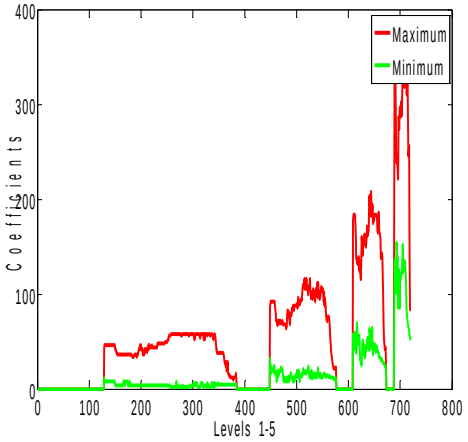


Fig. 7 Approximation at all five levels of decomposition

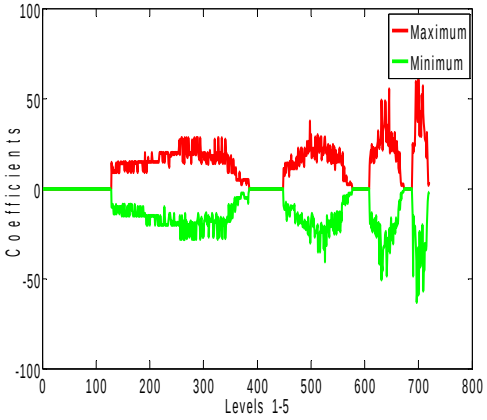


Fig. 8 Horizontal at all five levels of decomposition

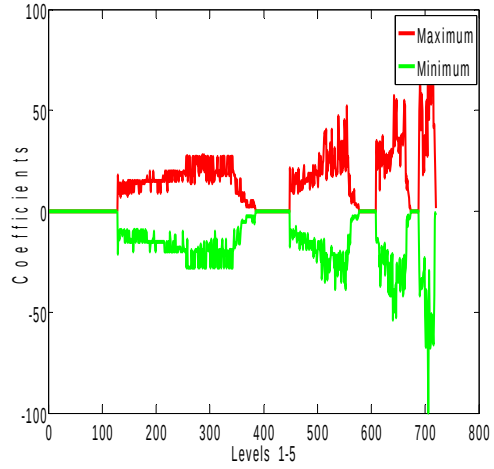


Fig. 9 Vertical at all five levels of decomposition

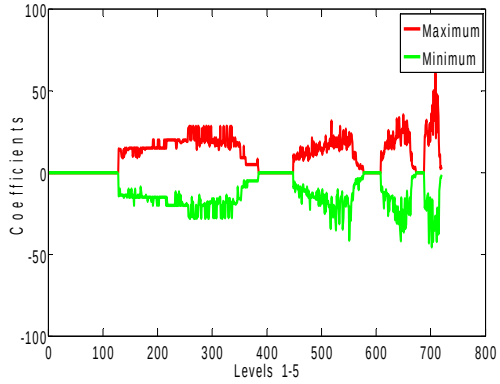


Fig. 10 Diagonal at all five levels of decomposition

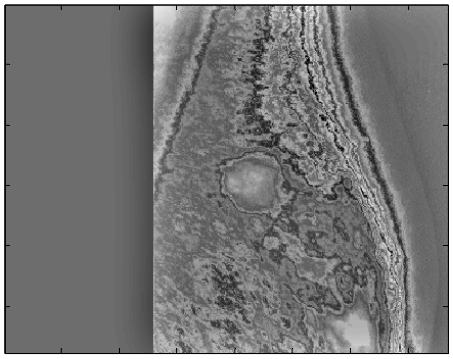


Fig.11 Presence of MC at the center

Training and testing the RBF

Training patterns were formed from sample mammogram under consideration. One hundred patterns were used for training the RBF. The number

.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

of input features used is 5 in the input layer of RBF ANN and the number of targets used is 1 in the output layer of ANN.

V. CONCLUSIONS

In this paper, Wavelets, radial basis function algorithm, and their combinations have been used on mammographic image analysis (MIAS) mammogram database. Radial basis function learns the patterns with one iteration. The performance of the RBF in identifying the MC depends upon the number centers used for training the RBF.

REFERENCES

[1]. Ackerman L. V., Mucciardi A. N., Gose E.E., and Alcorn F. S., 1973, "Classification of benign and malignant breast tumors on the basis of 36 radiographic properties", Cancer, Vol.31, pp.342–352.

[2]. Bovis K., Singh S.,. Fieldsend J, Pinder C., 2000, Identification of masses in digital mammograms with MLP and RBF nets, in: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks Com, pp.342–347.

[3]. Chang RF, WU WJ, Moon WK, Chen DR, 2005, Automatic Ultrasound Segmentation and Morphology based Diagnosis of Solid Breast Tumors. Breast Cancer Research and Treatment; Vol.89, No.2, pp.179–185.

[4]. Cheng HD, Cai X, Chen X, Hu L, Lou X., 2003, Computer-aided detection and classification of microcalcifications in mammograms: a survey. Pattern Recognition, Vol.36, pp.2967–91.

[5]. Chunekar, V.N., and Ambulgekar, H.P., 2009, Approach of Neural Network to Diagnose Breast Cancer on Three Different Data Set, Proceedings Advances in Recent Technologies in Communication and Computing 2009 (ARTcom-2009), IEEE, Kottayam. Pp.893-895.

[6]. Dehghan, F., and Abrishami-Moghaddam, H., 2008, Comparison of SVM and Neural Network classifiers in Automatic Detection of Clustered Microcalcifications in Digitized Mammograms. Proceedings 7th International Conference on Machine learning and Cybernetics 2008, ICMLA-2008, IEEE, Kunming., vol.2, pp.756-761.

[7]. Gholamali, R., Jamarani, S., 2005, Detecting Microcalcification Clusters Digital Mammograms Using Combination of Wavelet and Neural Network. In Proceedings of the Computer Graphics, Imaging and Vision: New Trends CGIV, pp.197-201.

[8]. Jamarani, SMH, Rezai-rad, G., Behnam, H , 2005, A Novel Method for Breast Cancer Prognosis Using Wavelet Packet Based Neural Network. In Proceedings of the IEEE 27th Annual International Conference on Engineering in Medicine and Biology Society, pp.3414–3417.

[9]. Jiang Y., Nishikawa R.M., Wolverton D.E., Metz C.E.,. Giger M.L, Schmidt R.A., 1996, Malignant and benign clustered microcalcifications: automated feature analysis and classification, Radiology, Vol.198, pp.671-678.

[10]. Leichter I., Lederman R., Buchbinder S., Bamberger P., Novak B., Fields S., 2000, Optimizing parameters for computer-aided diagnosis of microcalcifications at mammography, Acad. Radiol., Vol.7, pp.406-412.

[11]. Liao, R., Wan, T., and Qin, Z., 2011, Classification of Benign and Malignant Breast Tumors in Ultrasound Images Based on Multiple Sonographic and Textural Features, Proceedings International Conference on Intelligent Human-Machine Systems and Cybernetics 2011 (IHMSC-2011), IEEE, Hangzhou. pp.71-74.

[12]. Madabhushi A., and Metaxas D., 2003, Combining low-, high-level and Empirical Domain Knowledge for Automated Segmentation of Ultrasonic Breast Lesions, IEEE Transactions Medical Imaging, Vol.22, No.2, pp: 155–169.

[13]. Mahjabeen Mirza Beg, Monika Jain, 2012, An Analysis Of The Methods Employed For Breast Cancer Diagnosis, International Journal of Research in Computer Science ISSN 2249-8265, Vol.2, Issue 3, pp.25-29.

[14]. O'Neill T.J., Penm J., Penm J., 2007, A subset polynomial neural networks approach for breast cancer diagnosis, International Journal of Electronic Healthcare. Vol.3, Issue 3, pp.293-302.

[15]. Dhawan P., Chitre Y., and Kaiser-Bonasso C., 1996, Analysis of mammographic microcalcifications using gray-level image structure features, IEEE Trans. Med. Imaging, Vol.15, pp.246-259.

[16]. Papadopoulos A., Fotiadis D.I., Likas A., 2005, Characterization of Clustered Microcalcifications in Digitized Mammograms Using Neural Networks and Support Vector Machines. Artificial Intelligence in Medicine, Vol.34, pp.141-150.

[17]. .

[18]. Sampat, M.P., Markey, M.K., and Bovik, A.C., 2005, Computer-aided detection and diagnosis in mammography, in Handbook of Image and Video Processing, 2nd ed., New York: Academic Press. pp.1195-1217.

[19]. Strickland, R., Hahn, 1996, Wavelet Transforms for Detecting Microcalcifications in Mammograms. IEEE Transactions on Medical Imaging, Vol.15, No.2, pp.218-229.

[20]. Tsujii O., Freedman M.T and. Mun S.K, 1999, Classification of microcalcifications in digital mammograms using trend-oriented radial basis function neural network, Pattern Recognition, Vol.32, pp.891-903.

[21]. Veldkamp W.J.H., Karssemeijer N.. Otten J.D.M and Hendriks J.H.C.L., 2000, Automated classification of clustered microcalcifications into malignant and benign types, Med. Phys., Vol.27, pp.2600-2608.

[22]. Wang, T., Karayiannis, N., 1998, Detection of Microcalcification in Digital mammograms Using Wavelets, IEEE Transactions on Medical Imaging. Vol.17, No.4, pp.498-509.

SHAJI B, received the Master of Science in Computer Science from PG and Research Department of Computer Science & Applications, D.G.Vaishnav College, University of Madras in 2002 and M.Phil Computer Science from Alagappa University, Karaikudi, Tamil Nadu, India in 2005. Currently he is pursuing Ph.D in Computer Science from School of Computing Sciences, VELS Institute of Science Technology & Advanced Studies, Chennai, Tamil Nadu, India. He is working as Assistant Professor,

.

	Department of Computer Science, Asan Memorial College of Arts & Science, Chennai, Tamil Nadu, India since 2006. His area of interests includes Medical Image Processing, Artificial Neural Networks and Software Engineering.
	Dr.S.Purushothaman completed his PhD from Indian Institute of Technology Madras, India in 1995. He has 129 publications to his credit. He has 19 years of teaching experience. Presently he is working as Professor in PET Engineering College, India
	R.Rajeswari completed MSc Information Technology from Bharathidasan university, Tiruchirappalli and M.Phil Computer Science from Alagappa University, Karaikudi, Tamilnadu, India. She is currently pursuing PhD in Mother Teresa Women's University. Her area of interest is Intelligent Computing

.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IMPLEMENTATION OF INTRUSION DETECTION USING BPARBF NEURAL NETWORKS

[1]Kalpana Y., [2]Purushothaman S., and [3]Rajeswari R.,

[1]Kalpana Y., Research Scholar, VELS University, Pallavaram, Chennai, India-600117	[2]Dr.Purushothaman S., Professor, PET Engineering College, Vallioor, INDIA-627117.	[3]Rajeswari R., Research scholar, Mother Teresa Women's University, Kodaikanal-624102, INDIA.

Abstract: **Intrusion detection is one of core technologies of computer security. It is required to protect the security of computer network systems. Due to the expansion of high-speed Internet access, the need for secure and reliable networks has become more critical. The sophistication of network attacks, as well as their severity, has also increased recently. This paper focuses on two classification types: a single class (normal, or attack), and a multi class (normal, DoS, PRB, R2L, U2R), where the category of attack is detected by the combination of Back Propagation neural network (BPA) and radial basis function (RBF) Neural Networks. Most of existing IDs use all features in the network packet to look for known intrusive patterns. A well-defined feature extraction algorithm makes the classification process more effective and efficient. The Feature extraction step aims at representing patterns in a feature space where the attack patterns are attained. In this paper, a combination of BPA neural network along with RBF networks are used s for detecting intrusions. Tests are done on KDD-99 data set.**

Keywords: network intrusion detection, kdd-99 datasets, BPARABF neural networks

[I]. INTRODUCTION

With the tremendous growth of network-based services and sensitive information on networks, network security is getting more and more importance than ever. Intrusion [Alireza Osareh and Bita Shadgar, 2008] poses a serious security risk in a network environment. Intrusions [Asmaa Shaker Ashoor and Sharad Gore, 2011] are in many forms: attackers accessing a system through the Internet or insider attackers; authorized users attempting to gain and misuse non-authorized privileges. Intrusions are any set of actions that threaten the integrity, availability, or confidentiality of a network resource. Intrusion detection [Tich Phuoc Tran, et al, 2009] is the process of monitoring the events occurring in a computer system or network [Aida O. Ali, et al, 2010] and analyzing them for signs of intrusions.

Classification of Attack Detection

Attack/Invasion detection: It tries to detect unauthorized access by outsiders.
Misuse Detection: It tries to detect misuse by insiders, e.g., users who try to access services on the internet by passing security directives. Misuse detection uses a prior knowledge on intrusions and tries to detect attacks based on specific patterns of known attacks.
Anomaly Detection: It tries to detect abnormal states within a network.
Host Intrusion Detection System (HIDS): The HIDS works on information available on a system. It easily detects attacks by insiders as modification of files, illegal access to files and installation of Trojans.
Network Intrusion Detection System (NIDS): NIDS works on information provided by the network [Bahrololum, et al, 2009] mainly packets sniffed from the network layer. It uses protocol decoding, heuristical analysis and statistical anomaly analysis. NIDS detects DoS [Samaneh Rastegari, et al, 2009] with buffer overflow attacks, invalid packets, attacks on application layer and spoofing attacks.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[II]. RELATED WORKS

Zhang, et al, 2005, proposed a hierarchical IDS frameworks using RBF to detect both anomaly and misuse detection. A serial hierarchical IDS identifies misuse detection accurately and identifies anomaly detection adaptively. The purpose of parallel hierarchical IDS is to improve the performance of serial hierarchical IDS. Both the systems train themselves for new types of attacks automatically and detect intrusions real-time.

Meera Gandhi et al, 2009, propose a Polynomial Discriminant Radial Basis Function (PRBF) for intrusion detection to achieve robustness and flexibility. Based on several models with different measures, PRBF makes the final decision of whether current behavior is abnormal or not. Experimental results with some real KDD data show that the proposed fusion produces a viable intrusion detection system.

Ray-I Chang et al., 2007 proposed a learning methodology towards developing a novel intrusion detection system (IDS) by BPN with sample-query and attribute-query. The proposed method is tested by a benchmark intrusion dataset to verify its feasibility and effectiveness. Results showed that choosing attributes and samples will not only have impact on the performance, but also on the overall execution efficiency.

Ahmed Fares, et al., 2011, proposed two engines to identify intrusion, the first engine is the back propagation neural network intrusion detection system (BPNNIDS) and the second engine is the RBF neural network intrusion detection system and classify the attacks as two classification types: a single class (normal, or attack), and a multi class (normal, DoS, PRB, R2L, U2R). The model is tested against traditional and other machine learning algorithms using a common dataset: the DARPA 98 KDD99 benchmark dataset from International Knowledge

[III]. MATERIALS AND METHODOLOGIES

A. KDD CUP 1999 DATASET DESCRIPTION

The KDD Cup 1999 dataset has been used for the evaluation of anomaly detection methods. The KDD Cup 1999 contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type.

Data Collection: KDD Cup 1999 dataset has the different types of attacks: back, buffer_overflow, ftp_write, guess_passwd, imap, ipsweep, land, loadmodule, multihop, neptune, nmap, normal, perl, phf, pod, portsweep, rootkit, satan, smurf, spy, teardrop, warezclient, warezmaster. These attacks can be divided into 4 groups.

The Table 1 show the list of attacks in category wise:

Table 1 List of attacks

DoS	R2L	U2R	Probe
back	ftp_write	buffer_overflow	ipsweep
land	guess_passwd	loadmodule	nmap
neptune	imap	perl	portsweep
pod	multihop	rootkit	satan
smurf	phf		
teardrop	spy		
	warezclient		
	warezmaster		

Table 2 Sample KDD data

S. No.	KDD patterns
1	0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00, 1.00,0.00,0.00,255,254,1.00,0.01,0.00,0.00,0.00,0.00,0.00,0.00,normal.
2	0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00, 1.00,0.00,0.00,255,254,1.00,0.01,0.00,0.00,0.00,0.00,0.00,0.00,normal.
3	0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00, 1.00,0.00,0.00,255,254,1.00,0.01,0.00,0.00,0.00,0.00,0.00,0.00,normal.
4	0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00, 1.00,0.00,0.00,255,254,1.00,0.01,0.00,0.00,0.00,0.00,0.00,0.00,snmpgetattack.
5	0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00, 1.00,0.00,0.00,255,254,1.00,0.01,0.01,0.00,0.00,0.00,0.00,0.00,snmpgetattack.

B. Neural Network IDS

Two ANN networks are used: BPA and the RBF network.

Back Propagation neural network (BPNN)

The BPNN [Reyadh Shaker Naoum, et al, 2012, Meera Gandhi, et al, 2008] searches for weight values that minimize the total error of the network over a set of training examples. It consists of the repeated presentation of two passes: a forward pass

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

and a backward pass. In the forward pass, the network is activated and the error of each neuron of the output layer is computed. In the backward pass, the network error is used for updating the weights. This process is more complex, because hidden nodes are not directly linked to the error but are linked through the nodes of the next layer. Therefore, starting at the output layer, the error is propagated backwards through the network, layer by layer. This is achieved by recursively computing the local gradient of each neuron.
The training algorithm BPA are as follows:
1. Initialize the weights of the network randomly.
2. Present a training sample to the network where, in our case, each pattern consists of 41 features.
3. Compare the network's output to the desired output. Calculate the error for each output neuron.
4. For each neuron, calculate what the output should have been, and a scaling factor i.e. how much lower or higher the output must be adjusted to match the desired output. This is the local error.
5. Adjust the weights of each neuron to lower the local error.

$w_N = w_N + \Delta w_N$

with w_N computed using generalized delta rule

6. Repeat the process from step 3 on the neurons at the previous level.

Training Phase

A connection in the KDD-99 dataset is represented by 41 features. The features in columns 2, 3, and 4 in the KDD99 dataset are the protocol type, the service type, and the flag, respectively. The value of the protocol type may be tcp, udp, or icmp; the service type could be one of the different network services such as http and smtp; and the flag has 11 possible values such as SF or S2.

Weight Updation Methods

The neural network maps the input domains onto output domains. The inputs are packet parameters and the outputs are classification of attacks information. The combination of input and output constitutes a pattern. During training of ANN, the network learns the training patterns by a weight updating algorithm. The training of ANN is stopped when a desired performance index of the network is reached. The weights obtained at this stage are considered as final weights. During implementation of ANN for intrusion detection, the data coming from the network are transformed with the full weights obtained during the training of ANN. Every output of the network is checked. If the outputs are within the desired values detection is enabled.

Radial Basis Function network (RBFN)

A Radial basis function (RBF) network is a special type of neural network that uses a radial basis function as its activation function. A Radial Basis Function (RBF) neural network has an input layer, a hidden layer and an output layer. The neurons in the hidden layer contain radial basis transfer functions whose outputs are inversely proportional to the distance from the center of the neuron. In RBF networks, the outputs of the input layer are determined by calculating the distance between the network inputs and hidden layer centers. The second layer is the linear hidden layer and outputs of this layer are weighted forms of the input layer outputs. Each neuron of the hidden layer has a parameter vector called center. The RBF is applied to the distance to compute the weight for each neuron.
Centers are chosen randomly from the training set.

The following parameters are determined by the training process:
1. The number of neurons in the hidden layer.
2. The center of each hidden layer RBF function.
3. The radius of each RBF function in each dimension.
4. The weights applied to the RBF function outputs as they are passed to the summation layer.

The BPARBF methods have been used to train the networks.

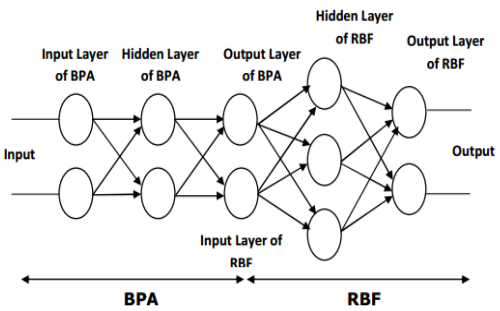


Fig.1 BPARBF Neural Network

The input layer provides elements of the input vector to all the hidden nodes. The nodes in the hidden layer holds the RBFs centers, computes the basis function to the Euclidean distance between the input vector and its centers. The nodes of hidden layer generates a scalar value, depends upon the centers it holds. The outputs of the hidden layer nodes are passed to the output layer via weighted connections. Each connection between the hidden

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

and output layers is weighted with the relevant coefficient. The node in the output layer sums its inputs to produce the network output.

Training RBF

Step 1: Initialize number of Inputs
Step 2: Create centers=Number of training patterns
Step 3: Calculate RBF as exp (-X) where X=(patterns-centers).
Step 4: Calculate Matrix as G=RBF and $A=G^T*G$.
Step 5: Calculate $B=A^{-1}$ and E=B * G*T.
Step 6: Calculate the final weight as F= (E*D) and store the final weights in a File.
Testing RBF
Step 1: Read output of BPA
Step 2: Calculate RBF as exp (-X) where X=(pattern-centers)
Step 3: Calculate Matrix as G=RBF
Step 4: Calculate Final value=Final weight * G.
Step 5: Classify the intrusion as an attack or normal.

[IV]. RESULTS AND DISCUSSIONS

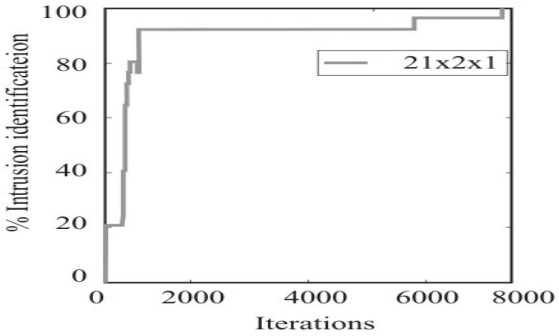


Fig2. Back propagation network for the intrusion detection

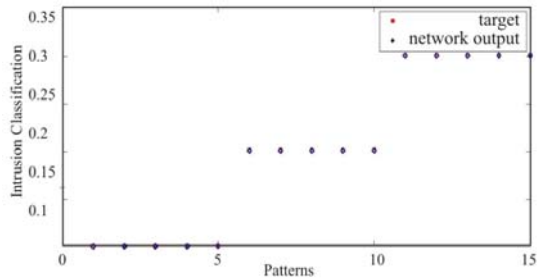


Fig.3 Radial Basis Function for intrusion detection

[V]. CONCLUSION

Current intrusion detection systems (IDS) examine data features to detect intrusion or misuse patterns. The purpose of this paper is to present combination of BPA with RBF for intrusion detection. Proper training of BPA and RBF results in detecting more number of intrusions.

REFERENCES

[1]. Ahmed H. Fares, Mohamed I. Sharawy, 2011, Intrusion Detection: Supervised Machine Learning, Journal of Computing Science and Engineering, Vol.5, No.4, pp.305-313.
[2]. Aida O. Ali, Ahmed saleh, Tamer Ramdan, 2010, Multilayer perceptrons networks for an Intelligent Adaptive intrusion detection system, IJCSNS International Journal of Computer Science and Network Security, Vol.10, No.2, pp.275-279.
[3]. Alireza Osareh, Bita Shadgar, 2008, Intrusion Detection in Computer Networks based on Machine Learning Algorithms, International Journal of Computer Science and Network Security, Vol.8, No.11, pp.15-23.
[4]. Asmaa Shaker Ashoor and Sharad Gore, 2011, Importance of Intrusion Detection System, International Journal of Scientific and Engineering Research, Vol.2, Issue 1, pp.1-4.
[5]. Bahrololum M., Salahi E., Khaleghi M., 2009, Anomaly Intrusion Detection Design Using Hybrid Of Unsupervised And Supervised Neural Network, International Journal of Computer Networks and Communications (IJCNC), Vol.1, No.2, pp.26-33.
[6]. Meera Gandhi, Srivatsava S.K., 2008, Application of Back propagation Algorithm in Intrusion Detection in Computer Networks, International Journal of Soft computing, Vol.3, No.4, pp.277-281.
[7]. Meera Gandhi, Srivatsa S.K, 2009, Polynomial Discriminant Radial Basis Function for intrusion detection, International Journal of Cryptography and Security, Vol.2, No.1, pp.25-32.
[8]. Ray-I Chang, Liang-Bin Lai, Wen-De Su, Jen-Chieh Wang, Jen-Shiang Kouh, 2007, Intrusion Detection by Back propagation Neural Networks with Sample-Query and Attribute-Query, International Journal of Computational Intelligence Research. Vol.3, No.1, pp.6-10.
[9]. Reyadh Shaker Naoum, Namh Abdula Abid, Zainab Namh Al-Sultani, 2012, An Enhanced Resilient Backpropagation Artificial Neural Network for Intrusion Detection System, International Journal of Computer Science and Network Security, Vol.12, No.3, pp.11-16.
[10]. Samaneh Rastegari, Iqbal Saripan M., Mohd Fadlee A., Rasid, 2009, Detection of Denial of Service Attacks against Domain Name System Using Neural Networks, IJCSI International Journal of Computer Science Issues, Vol.6, No.1, pp.23-27.
[11]. Tich Phuoc Tran, Longbing Cao, Dat Tran, Cuong Duc Nguyen, 2009, Novel Intrusion Detection using Probabilistic Neural Network and Adaptive Boosting, International Journal of Computer Science and Information Security, Vol.6, No.1, pp.83-92.
[12]. Zhang Chunlin, Ju Jiang, Mohamed Kamel, 2005, Intrusion detection using hierarchical neural networks, Pattern Recognition Letters 26, Vol.9, No.45, pp.779–791.

	Y.Kalpana has received her M.C.A and M.Phil. degrees from Bharathidasan university, India and currently pursuing her Ph.D degree in VELS University. She has 15 years of Teaching experience. She has presented 8 papers in National Conference and 1 paper in International conference. Her research interests include Network security and Data Mining.
	Dr.S.Purushothaman completed his PhD from Indian Institute of Technology Madras, India in 1995. He has 129 publications to his credit. He has 19 years of teaching experience. Presently he is working as Professor in PET college of Engineering, India
	R.Rajeswari completed MSc Information Technology from Bharathidasan university, Tiruchirappalli and M.Phil Computer Science from Alagappa University, Karaikudi, Tamilnadu, India. She is currently pursuing PhD in Mother Teresa Women's University. Her area of interest is Intelligent Computing

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IMPLEMENTATION OF DAUBAUCHI WAVELET WITH RADIAL BASIS FUNCTION AND FUZZY LOGIC IN IDENTIFYING FINGERPRINTS

[1]Guhan P.,[2]Purushothaman S., and [3]Rajeswari R.,

[1]Guhan P., Research Scholar, Department of MCA, VELS University, Chennai–600 117, India

[2]Dr.Purushothaman S.,Professor, PET Engineering College, Vallioor, India-627117,

[3]Rajeswari R., Research scholar, Mother Teresa Women's University, Kodaikanal-624102, India.

***Abstract*-This paper implements wavelet decomposition for extracting features of fingerprint images. These features are used to train the radial basis function neural network and Fuzzy logic for identifying fingerprints. Sample finger prints are taken from data base from the internet resource. The fingerprints are decomposed using daubauchi wavelet 1(db1) to 5 levels. The coefficients of approximation at the fifth level is used for calculating statistical features. These statistical features are used for training the RBF network and fuzzy logic. The performance comparisons of RBF and fuzzy logic are presented.**

***Keywords*- Fingerprint;Daubauchiwavelet, radial basis function, fuzzy logic.**

I. INTRODUCTION

Fingerprint image databases are characterized by their larger size. Distortions are very common in fingerprint images due to elasticity of the skin. Commonly used methods for taking fingerprint impressions involve applying a uniform ink on the finger and rolling the finger on the paper. This causes

1. over-inked areas of finger, which create smudgy areas in the images,
2. breaks in ridges, created by–under-inked areas,
3. the elastic nature of the skin can cause positional shifting, and
4. thenon-cooperative attitude of criminals also leads to smearing in parts of the fingerprint images.

Although inkless methods for taking fingerprint impressions are now available, these methods also suffer from the positional shifting caused by the skin elasticity. Thus, a substantial amount of research reported in the literature on fingerprint identification is devoted to image enhancement techniques.

Current approaches in pattern recognition to search and query large image databases, based upon the shape, texture and color are not directly applicable to fingerprint images. The contextual dependencies present in the images and the complex nature of two dimensional images make the representational issue very difficult. It is very difficult to find a universal content-based retrieval technique. For these reasons an invariant image representation of a fingerprint image[Islam, et al, 2010;Pokhriyal and Lehri, 2010] is still an open research issue.

The problems associated with fingerprint identification [Pankanti,et al, 2002] are very complex, and an inappropriate representation scheme can make it intractable. For the purpose of automating the process of fingerprint identification, a suitable representation of fingerprints is essential. But these representations do not guarantee exact matching because of the presence of noise or availability of a partial image. Hence, high level structural features, which can uniquely represent a fingerprint, are extracted from the image for the purpose of representation and matching.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

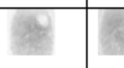
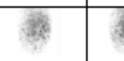
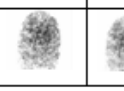
II. RELATED WORK

Fingerprint recognition[Azizun, et al, 2004, Tico et al, 2001] was formally accepted as a valid personal identification method and became a standard routine in forensics.Wavelet based features(WBF)[Thaiyalnayaki, et al, 2010] extracted from the channel impulse response (CIR) in conjunction with an artificial neural network (ANN). Honglie Wei and Danni liu, 2009, proposed a fingerprint matching technique based on three stages of matching which includes local orientation structure matching, local minutiae structure matching and global structure matching.Chengming wen, et al, 2009, have proposed an algorithm for one to one matching of minutiae points using motion coherence methods. The K-plot was used to describe local structure. Ujjal Kumar Bhowmik et al, 2009, proposed that smallest minimum sum of closest Euclidean distance (SMSCED) corresponding to the rotation angle to reduce the effect of non linear distortion. The overall minutiae patterns of the two fingerprints are compared by the SMSCED between two minutiae sets. Khuramand Shoab, 2009, proposed fingerprint matching using five neighbor of one single minutiae i.e., center minutiae. The special matching criteria incorporate fuzzy logic to select final minutiae for matching score calculation.Anil K. Jain, 2009, proposed algorithm to compare the latent fingerprint image with that of the stored in the template. From the latent fingerprint minutiae orientation field and quality map are extracted. Both level 1 and 2 features are employed in computing matching scores.Quantitative and qualitative scores are computed at each feature level. Xuzhou Li and Fei Yu, 2009, proposed fingerprint matching algorithm that uses minutiae centered circular regions. The circular regions constructed around minutiae are regarded as a secondary feature. The minutiae pair that has the higher degree of similarity than the threshold is selected as reference pair minutiae. Jian-De Zheng, et al, 2009, introduced fingerprint matching based on minutiae. The proposed algorithm uses a method of similar vector triangle. The ridge end points are considered as the reference points. Using the reference points the vector triangles are constructed. The fingerprint matching is performed by comparing the vector triangles.

III. MATERIALS AND METHODOLOGY

A sample database is presented for 10 people in Table 1. Each row presents 4 fingerprints of a person. Similarly, there are 10 rows showing 10 people.

Table 1 Sample fingerprint database

	Event 1	Event 2	Event 3	Event 4
Person 1				
Person 2				
Person 3				
Person 4				
Person 5				
Person 6				
Person 7				
Person 8				
Person 9				
Person 10				

A WAVELETS

The wavelet (WT) was developed as an alternative to the short time fourier transform (STFT). A wavelet is a waveform of effectively limited duration that has an average value of zero. Compare wavelets with sine waves, which are the basis of Fourier analysis. Sinusoids do not have limited duration, they extend from minus to plus infinity and where sinusoids are smooth and predictable, wavelets tend to be. Wavelet analysis is the breaking up of a signal into shifted and scaled versions of the original (or mother) wavelet. Mathematically, the process of Fourier analysis is represented by the Fourier transform: which is the sum over all time of the signal f(t) multiplied by a complex exponential. The results of the transform are the Fourier coefficients, which when multiplied by a sinusoid of frequency, yield the constituent sinusoidal components of the original signal. The continuous wavelet transform (CWT) is defined as the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet function. The result of the CWT is many wavelet coefficients C, which are a function of scale and position. Multiplying each coefficient by the appropriately scaled and shifted wavelet yields the constituent wavelets of the original signal.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

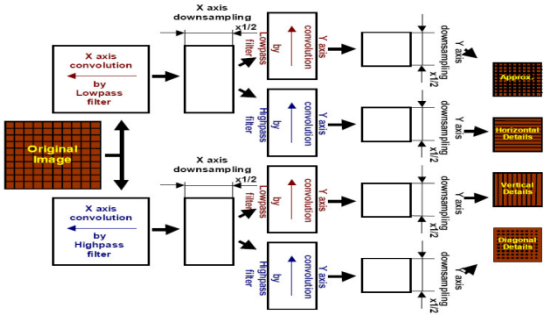


Fig.1. Decomposition using wavelet

(Courtesy:http://software.intel.com/ sites/ products/ documentation/ hpc/ipp/ippi/ippi_ch13/ ch13_Intro.html)

An image can be analyzed for various information by decomposing the image using wavelet of our choice. Decomposition operation applied to a source image produces four output images of equal size: approximation image, horizontal detail image, vertical detail image, and diagonal detail image.The flow of decomposition process is shown in Figure 1. Fingerprint image is given as input to the system and level 1 to level decompositions take place. Initially, Approximation, horizontal, vertical and diagonal matrices are obtained from the original image. Each matrix is ¼th size of the input image. In the level two and subsequent levels, Approximation matrix of the previous levels are used for subsequent decompositions.

These decomposition components have the following meaning:

1. The 'approximation' image is obtained by vertical and horizontal lowpass filtering.
2. The 'horizontal detail' image is obtained by vertical highpass and horizontal lowpass filtering.
3. The 'vertical detail' image is obtained by vertical lowpass and horizontal highpass filtering.
4. The 'diagonal detail' image is obtained by vertical and horizontal highpass filtering.

Proposed method

Step 1: Fingerprint image is decomposed using db1 to 5 levels.

Step 2: The coefficients of approximation at 5th level is used for training the RBF network and Fuzzy logic.

Step 3: At the end of training process, the final weights are stored in a file.

Step 4: During the testing process, the decomposition to 5th level using db1 and statistical feature extraction are done. The features are processed with final weights of RBF and Fuzzy logic to identify fingerprint.

Wavelet Features Extraction

The features are obtained from the Approximation and Details of the 5th level by using the following equations

V1=1/d ∑ (Approximation details)

Where d = Samples in a frame and V1 = Mean value of approximation

V2=1/d ∑ (Approximation or details

Where V2=Standard Deviation of approximation

V3=maximum (Approximation or details)

V4=minimum (Approximation or details)

V5=norm (Approximation or Details)^2

Where V5 = Energy value of frequency

B. RADIAL BASIS FUNCTION (RBF)

Radial basis function is a supervised neural network. The network has an input layer, hidden layer (RBF layer) and output layer. The features obtained from daubauchi wavelet decompositions are used as inputs for the network along with target values. The network (Figure 2) described is called an RBFNN, since each training data point is used as a basis center. The storage costs of an exact RBFNN can be enormous, especially when the training database is large.

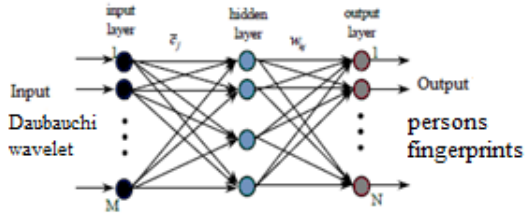


Fig.2. The Radial basis function neural network

Training RBF is done as follows,

Step 1: Finding distance between pattern and centers.

Step 2: Creating an RBF matrix whose size will be (np X cp). , where np= number of fingerprint patterns (50 fingerprint patterns X number of patterns) used for training and cp is number of centers which is equal to 50. The number of centers chosen should make the RBF network learn the maximum number of training patterns under consideration.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Step 3: Calculate final weights which are inverse of RBF matrix multiplied with Target values.

Step 4: During testing the performance of the RBF network, RBF values are formed from the features obtained from fingerprint image and processed with the final weights obtained during training. Based on the result obtained, the image is classified to particular fingerprint.

Training RBFfor identifying fingerprints

Step 1: Apply Radial Basis Function.

No. of Input = 5

No. of Patterns = 50

No. of Centers= 50

Calculate RBF as

RBF = exp (-X)

CalculateMatrix as

$G = RBF$

$A = G^T * G$

Calculate

$B = A^{-1}$

Calculate

$E = B * G^T$

Step 2: Calculate the Final Weight.

$F = E * D$

Step 3: Store the Final Weights in a File.

Testing RBFfor identifying fingerprints

Step 1:.Read the Input

Step 2: Read the final weights

Step 3 Calculate.

Numerals = F * E

Step 4: Check the output with the templates

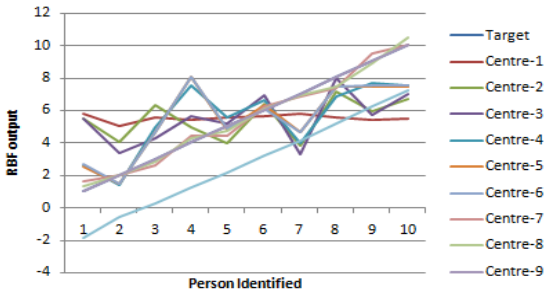


Fig 3 Performance of RBF

Figure 3 presents number of persons' fingerprints and RBF network estimation. All the 10 fingerprints are correctly identified only when the RBF center is 9. When the RBF centers are less or more than 9, then fingerprint identification performance comes down

C.Fuzzy logic

Fuzzy Logic (FL) is a multi valued logic that allows intermediate values to be defined between conventional evaluations like true/false, yes/no, high/low. Fuzzy systems are an alternative to traditional notions of set membership and logic.

The training and testing fuzzy logic is to map the input pattern with target output data. For this, the inbuilt function has to prepare membership table and finally a set of number is stored. During testing, the membership function is used to test the pattern.

Training Fuzzy logic for identifying fingerprints

Step 1: Read the statistical features of the wavelet coefficients and its target value.

Step 2: Create Fuzzy membership function.

Step 3: Create clustering using K-Means algorithm.

Step 4: Process with target values.

Step 5: Obtain final weights.

Testing Fuzzy logic for identifying fingerprints

Step 1: Input a pattern (statistical features of the wavelet coefficients).

Step 2: Process with Fuzzy membership function.

Step 5: Find the cluster to which the pattern belongs.

Step 4: Obtain estimated target values.

Step 5: Classify the fingerprint

RADII specifies the range of influence of the cluster center for each input and output dimension, assuming the data falls within a unit hyperbox (range [0 1]). Specifying a smaller cluster radius will usually yield more, smaller clusters in the data, and hence more rules. When RADII is a scalar it is applied to all input and output dimensions.

IV. RESULTS AND DISCUSSION

The coefficient values are presented 'approximation' (Figure 4), 'horizontal' (Figure 5), 'vertical' (Figure 6) and 'details (Figure 7) at 5th level of decomposition using 'db1' wavelet. Figure 8 presents fingerprints at all 5 levels for the fingerprint of person 1 with event 1.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

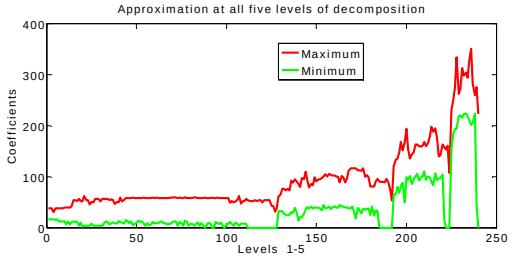


Fig. 4. Approximation at all 5 levels

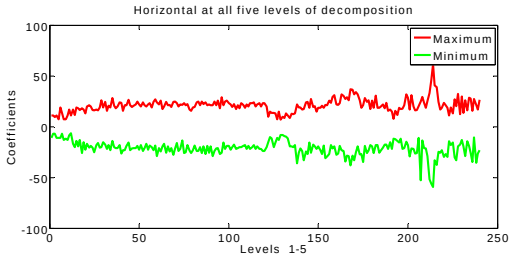


Fig. 5. Horizontal at all 5 levels

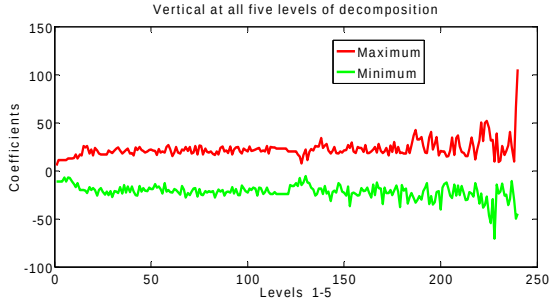


Fig. 6. Vertical at all 5 levels

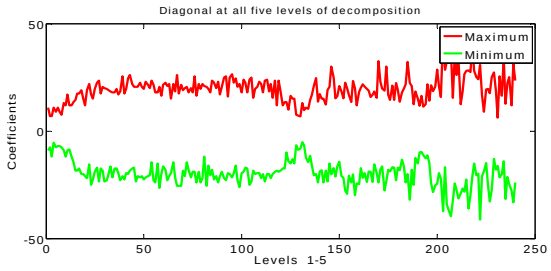


Fig. 7. Details at all 5 levels

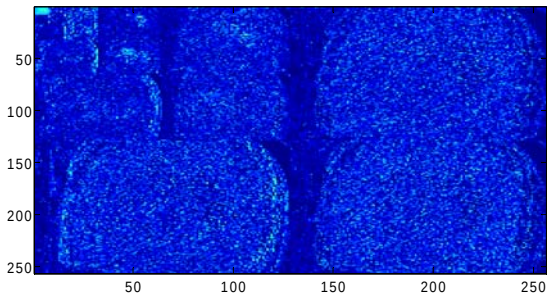


Fig. 8. Fingerprints shown at 5 levels of decompositions

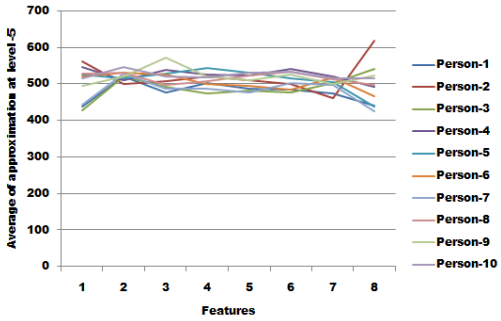


Fig.9. Statistical feature of Approximation at Level-5 decompositionof fingerprint images of 10 People

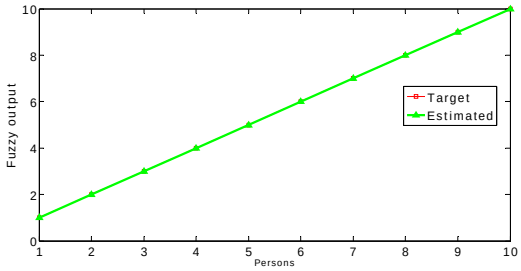


Fig.10. Performance of Fuzzy logic

Figure 10 presents number of persons' fingerprints and Fuzzy logic estimation. In all the 10 fingerprints, the estimation is 100%. The performance of Fuzzy logic may change, if the number of fingerprints increase.

V. CONCLUSION

This paper presents the implementation of radial basis neural network and fuzzy logic for identifying fingerprints. The features of the fingerprint images are obtained by using wavelet decomposition. The fingerprints have been collected from the existing available internet database. The proposed algorithms are able to identify the fingerprints.

REFERENCES

[1]. Anil K. Jain, JianjiangFeng, Abhishek Nagar and KarthikNandakumar, 2008, On Matching Latent Fingerprints, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp.1-8.

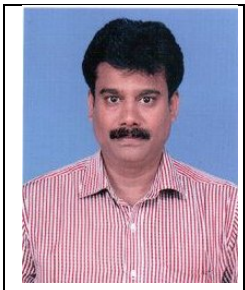
[2]. Azizun W. Adnan, Siang L. T., Hitam S., 2004 Fingerprint recognition in wavelet domain, Journal Teknologi, 41(D), pp.25-42.

[3]. Chengming Wen, TiandeGuo and Shuguang Wang, 2009, Fingerprint Feature-point Matching Based on Motion Coherence, Second International Conference on Future Information technology and Management Engineering, pp.226-229.

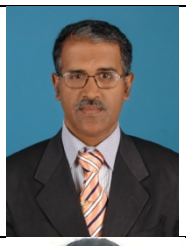
[4]. Honglie Wei and Danni Liu, 2009, A Multistage Fingerprint Matching Algorithm,

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Proceedings of the IEEE International Conference on Automation and Logistics, pp.197-199.
[5]. Islam Md. I, Begum N., Alam M., and Amin M.R, 2010, Fingerprint Detection Using Canny Filter and DWT, a New Approach, Journal of Information Processing Systems, , Vol.6, No.4, pp.511-520.
[6]. Jian-De Zheng, Yuan Gao and Ming-Zhi Zhang, 2009, Fingerprint Matching Algorithm Based on Similar Vector Triangle, Second International Congress on Image and Signal Processing, pp.1-6.
[7]. KhurramYasinQureshi and Shoab A. Khan, 2009, Effectiveness of Assigning Confidence Levels to Classifiers and a Novel Feature in Fingerprint Matching, IEEE International Conference on Systems, Man, and Cybernetics, pp.2181-2185.
[8]. Pankanti, S. Prabhakar, and Jain A.K., 2002, On the individuality of fingerprints, IEEE Trans. Pattern Anal. Mach. Intell., Vol.24, No.8, pp.1010–1025.
[9]. Pokhriyal A., and Lehri S., 2010, A New Method of Fingerprint Authentication Using 2D Wavelets, Journal of Theoretical and Applied Information Technology, Vol.13, No.2, pp.131-138.
[10]. Thaiyalnayaki K., Karim S.A., Parmar P.V., 2010, Finger print Recogntion Using Discrete Wavelet Transform", International Journal of Computer Applications, Vol.1,No.24, pp.96-100.
[11]. Tico M., Kuosmanen P., and. Saarinen J., 2001, Wavelet domain features for fingerprint recognition, Electronics Letters, Vol.37, No.1. pp.21-22.
[12]. Ujjal Kumar Bhowmik, AshkanAshrafi and Reza R. Adhami, 2009, A Fingerprint Verification Algorithm Using the Smallest Minimum Sum of Closest Euclidean Distance, IEEE International Conference on Electrical, Communications and Computers, pp.90-95.
[13]. Xuzhou Li and Fei Yu, 2009, A New Fingerprint Matching Algorithm Based on Minutiae, IEEE International Conference on Communications Technology and Applications, pp.869-873.



P.GUHAN CompletedM.C.A.,atShanmugaCollegeofEngineering,Thanjavur,andM.Phil.,inComputerScienceatPeriyarUniversity,Salem.Hehas8Publicationstohiscredit. Hehas13yearsofTeachingandResearchexperience.PresentlyheisworkingasAssistantProfessorinDepartmentofM.C.A.,JAYACOLLEGEOFARTS&SCIENCE,ChennaiandcurrentlypursuingPhD.,inVELSUniversity,Chennai.



Dr.S.Purushothaman completed his PhD from Indian Institute of Technology Madras, India in 1995. He has 129 publications to his credit. He has 19 years of teaching experience. Presently he is working as Professor in PET Engineering College, India

R.Rajeswari completed MSc Information Technology from Bharathidasan university, Tiruchirappalli and M.Phil Computer Science from Alagappa University, Karaikudi, Tamilnadu, India. She is currently pursuing PhD in Mother Teresa Women's University. Her area of interest is Intelligent Computing

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IMPLEMENTATION OF HIDDEN MARKOV MODEL AND COUNTER PROPAGATION NEURAL NETWORK FOR IDENTIFYING HUMAN WALKING ACTION

[1]Sripriya P., [2]Purushothaman S., and [3]Rajeswari R.,

[1]Research Scholar,
Department of MCA, VELS University, Pallavaram, Chennai, India-600117, Email:

[2]Dr.Purushothaman S,
Professor, PET Engineering College, Vallioor, India-627117,

[3]Rajeswari R,
Research scholar,
Mother Teresa Women's University, Kodaikanal, India-624101.

***Abstract*-This paper presents the combined implementation of counter propagation network (CPN) along with hidden Markov model (HMM) for human activity recognition. Many methods are in use. However, there is increase in unwanted human activity in the public to achieve gainsay without any hard work. Surveillance cameras are installed in the crowded area in major metropolitan cities in various countries. Sophisticated algorithms are required to identify human walking style to monitor any unwanted behavior that would lead to suspicion. This paper presents the importance of CPN to identify the human GAIT.**

Keywords: GAIT; human walking action; counter propagation network; hidden Markov model.

I. INTRODUCTION

Recognizing human activities from video is one of the most promising applications of computer vision. In recent years, this problem has caught the attention of researchers from industry, academia, security agencies, consumer agencies and the general populace too. One of the earliest investigations into the nature of human motion was conducted by the contemporary photographers Etienne Jules Marey and Eadweard Muybridge in the 1850s who photographed moving subjects and revealed several interesting and artistic aspects involved in human and animal locomotion. The classic Moving Light Display (MLD) experiment of (Johansson, 1973) provided a great impetus to the study and analysis of human motion perception in the field of neuroscience. This paved the way for mathematical modeling of human action and automatic recognition, which naturally fall into the purview of computer vision and pattern recognition.

Given a sequence of images with one or more persons performing an activity, can a system be designed that can automatically recognize what activity is being or was performed? As simple as the question seems, the solution has been that much harder to find. Aggarwal and Cai, 1999, discuss three important sub-problems that together form a complete action recognition system extraction of human body structure from images, tracking across frames and action recognition.

Cedras and Shah, 1995, present a survey on motion based approaches to recognition as opposed to structure based approaches. They argued that motion is a more important cue for action recognition than the structure of the human body. Gavrila, 1999, presented a survey focused mainly on tracking of hands and humans via 2D or 3D models and a discussion of action recognition techniques. Moeslund et al, 2006, presented a survey of problems and approaches in human motion capture including human model initialization, tracking, pose estimation and activity recognition.

II. RELATED WORK

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The task of recognizing people by the way they walk is an instance of the more general problem of recognition of humans from gesture or activity. A closer look at the relation between the problems of activity recognition and activity-specific person identification is considered. A good review of the state of the art in activity recognition can be found, (Aggarwal et al., 1999). For human activity or behavior recognition most efforts have used HMM-based approaches (Starner et al., 1998; Wilson et al, 1998; Yamato et al, 1995) as opposed to template matching which is sensitive to noise and the variations in the movement duration. Discrete HMMs are used to recognize different tennis strokes.

A parametric continuous HMM has been applied for activity recognition. All these approaches involve picking a lower dimensional feature vector from an image and using these to train an HMM. The trajectories corresponding to distinct activities will be far apart in the vector space of the features. Hence, with a small degradation in performance, it is possible to replace the continuous approaches by building a codeword set through k-means clustering over the set of the lower dimensional observation vector space and using a discrete HMM approach. The scenario is very different in the problem of recognition of humans from activity. Primarily, there is considerable similarity in the way people perform an activity. Hence, feature trajectories corresponding to different individuals performing the same activity tend to be much closer to one another as compared to feature trajectories corresponding to distinct activities. The afore mentioned activity recognition approaches, if directly applied to human identification using gait will almost certainly fail in the presence of noise and structurally similar individuals in the database.

Huang et al., 1999, use optical flow to derive the motion image sequence corresponding to a gait cycle. The approach is sensitive to optical flow computation. It does not address the issue of phase in a GAIT cycle. In another approach, (Cunado et al, 1995) extract gait signature by fitting the movement of the thighs to an articulated pendulum like motion model. The idea is somewhat similar to the work by (Murray et al, 1964), who modeled the hip rotation angle as a simple pendulum, the motion of which was approximately described by simple harmonic motion.

Locating accurately the thigh in real image sequences can be very difficult. Little and Boyd, 1998, extracted frequency and phase features from moments of the motion image derived from optical flow to recognize different people by their gait. As expected, the method is quite sensitive to the feature extraction process. Bobick and Johnson, 2001, used static features for recognition of humans using GAIT. Murase and Sakai, 1996, have also proposed a template matching method.

III. MATERIALS AND METHODOLOGY

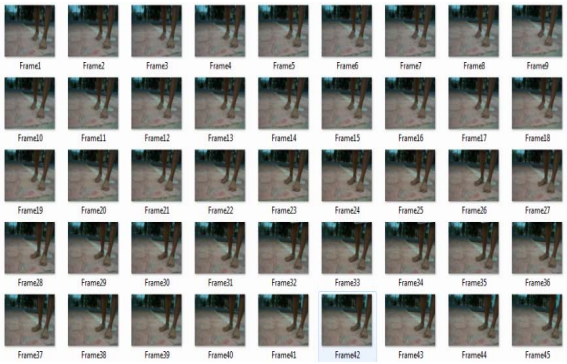


Fig 1 Keyframes for walking

A. HIDDEN MARKOV MODEL

A HMM is a statistical model where the system modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters, from the observable parameters, based on assumptions. The extracted model parameters can be used to perform further analysis to recognize human activity. A HMM adds outputs: each state has a probability distribution, (Sasi, et al, 2011).

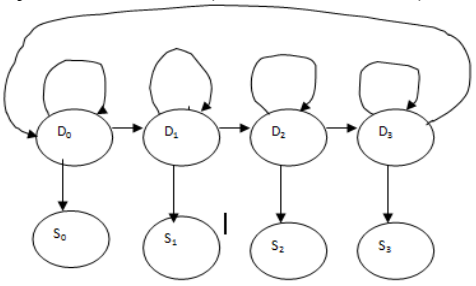


Fig. 2 Cyclic Left Right Hidden Markov Model

A Hidden Markov Model is a five-tuple consisting of set of states is given in equation (1).

$$S= \{s_1,s_2,...,s_n\} \quad (1)$$

The initial probability distribution, $\pi(s_i)$ = probability of s_i being the first state in a sequence.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The matrix of state transition probabilities, $A= (a_{ij})$ where a_{ij} is the probability of state s_j following s_i .

The set of emission probability distributions / densities is given in equation (2).

$$B= \{b_1, b_2, b_n\} \quad (2)$$

where

$b_i(x)$ is the probability of observing x.

When the system is in state s_i, the observable feature space can be discrete which is given in equation (3).

$$V=\{x_1, x_2,\ldots.., x_v\} \quad (3)$$

Let, $\lambda = \{A, B, b_i\}$ denote the parameters for a given HMM with fixed 'S' and 'V'. The left-right HMM Model is used for the walking cycle where a state can only be reached by itself or by a single other state. The states are hidden and one can only observe a sequence of observations generated from a sequence of states. In this work, skeletal joint configurations of a person are observed. Figure 2 shows the left-right HMM.

In Figure 2, the 'D' nodes represent the hidden states and the 'S' nodes give the observation from these states. A HMM is parameterized by, transition probabilities a_{ij}, $1 \leq i, j \leq N$, which are the probabilities of transitioning from state D_i to state D_j in the successive time. Initial probabilistic, a_{ij}, $1 \leq i \leq N$, which are the probabilities of a sequence starting in any state 'D_i' and observation probabilities '$b_i(s)$', $1 \leq i \leq N$, which are the probabilities of observing 'x' given being in state 'D_i'.

These parameters can be summarized by $\gamma = (A, B, b_i)$ where 'b_i' is a vector of probabilities and 'A' is matrix of probabilities. For this system, the continuous Gaussian observation is modeling and so 'B' will be replaced with a Gaussian Probability density function, $N(s, \mu, \sum)$ where μ and $\sum$ are the mean and co-variances of the distributions.

The goal is to automatically learn the parameter, $\gamma= (A, N)$ from motion captured data of walking sequence. Each person's walk will be modeled by one identically structured HMM. The unique parameter learnt will be used to recognize people from new observation sequences. By performing random walk on the HMM, new walking motion for each person can produce.

After extracting the necessary feature vector, the continuous K-means clustering algorithm is used. The continuous K-means algorithm identifies the matrix data and represents the data which is clustered. This clustered data gives the input to the initializer block where the parameters of HMM $\gamma = (A, B)$ are initialized.

The parameter $\gamma = (A, N)$ are initialized in the HMM. The joint contributions of walking sequences as continuous observation from discrete states are modeled here. These continuous observations using $b_i(s)=N(s; \mu_i, \sigma_i)$, $1 \leq i \leq N$ are modeled, where '$b_i(s)$' model the probability density function of observation 's' generated from state 'i' as a Gaussian. The 4 partitions (N=4) of the feature vector found by K-means are used to initialize the Gaussian parameter 'μ', 'σ'. The identified cluster means corresponds directly to the mean vector 'μ_i'. The form of 'σ_i' is spherical where the covariance is set to be the squared distance to the next closest mean vector. The probability to the prior of the HMM is distributed. This means that it is equally likely to begin in any state of our HMM as assumed. The transition matrix 'A', specifying the probability of transitioning from any state to any other state is initialized based on frequency counts from the feature vector data. Each feature vector corresponds to a single frame in a sequence. From this, and the partition information found in the last section, it is found how many times, from one state to another state in consecutive frames that is

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the element a_{ij} ($i \leq 1$, $j \leq N$) in 'A' can be transmitted, estimate the probability $P(D_{i,t+1} / D_{i,t})$ of transitioning from state D_i to state D_j in consecutive frames by counting the number of times in the sequence that will transit from state D_i to D_j divided by the total number of time in the state D_i.

General steps adopted in implementing the coding of HMM

a. The proposed system attempt to recognize people by modeling each individual's GAIT using a HMM. The HMM is a good choice for modeling a walk cycle because it can model sequential processes.

b. The HMM can be built by converting motion capture data from .amc file format to Matlab matrices (Converter).

c. Using feature vectors from that data to produce cluster sets (Clusterer), initializing the model with appropriate values based on the cluster set (Initializer) and then improving the model parameters to best represent the original data (Trainer).

d. Then the HMM for each person can make use of the model in two ways: identifying individuals from new unseen data (Identifier) or from training data so as to test the model's accuracy (Tester), and generating new pieces of motion (Generator).

B.COUNTER PROPAGATION NEURAL (CPN) NETWORK

CPN was developed by Robert Hecht-Nielsen as a means to combine an unsupervised Kohonen layer with a teachable output layer known as Grossberg layer, by James et al (1991), Athanasios et al. (2004). The operation of this network type is very similar to that of the Learning Vector Quantization (LVQ) network in that the middle (Kohonen) layer acts as an adaptive look-up table.

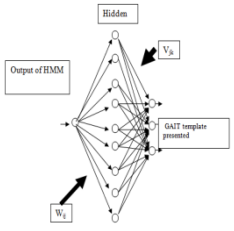


Fig. 3 Training the CPN

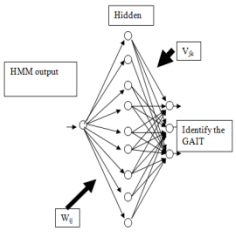


Fig. 4 Testing the CPN

The Figure 5 gives the flowchart of the CPN. From this figure, it is clear that the counter-propagation network is composed of three layers: an input layer that reads input patterns from the training set and forwards them to the network, a hidden layer that works in a competitive fashion and associates each input pattern with one of the hidden units, and the output layer which is trained via a teaching algorithm that tries to minimize the mean square error (MSE) between the actual network output and the desired output associated with the current input vector. In some cases, a fourth layer is used to normalize the input vectors but this normalization can be easily performed by the application before these vectors are sent to the Kohonen layer.

Regarding the training process of the counter-propagation network, it can be described as a two-stage procedure: in the first stage, the process updates the weights of the synapses between the input and the Kohonen layer, while in the second stage the weights of the synapses between the Kohonen and the Grossberg layer are updated.

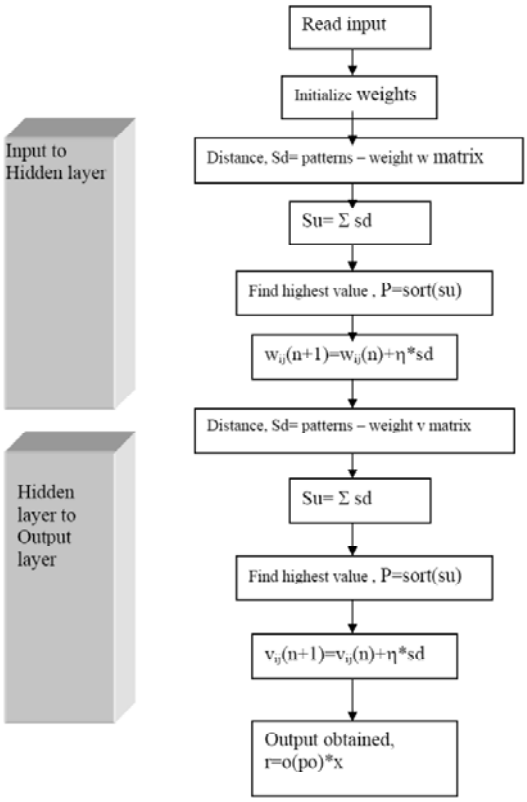


Fig. 5 Flow chart for Counter propagation network

Training of the weights from the input to the hidden nodes

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Step 1: The synaptic weights of the network between the input and the Kohonen layer are set to small random values in the interval [0, 1].

Step 2: A vector pair (x, y) of the training set, is selected at random.

Step 3: The input vector 'x' of the selected training pattern is normalized.

Step 4: The normalized input vector is sent to the network.

Step 5: In the hidden competitive layer, the distance between the weight vector and the current input vector is calculated for each hidden neuron 'j' according to the equation (4)

$$D_j = \sqrt{\sum_{i=1}^{K} (x_j - w_{ij})^2} \quad \textbf{(4)}$$

where

K is the number of the hidden neurons and

W_{ij} is the weight of the synapse that joins the i^{th} neuron of the input layer with the j^{th} neuron of the Kohonen layer.

Step 6: The winner neuron 'W' of the Kohonen layer is identified as the neuron with the minimum distance value 'D_j'

Step 7: The synaptic weights between the winner neuron 'W' and all neurons of the input layer are adjusted according to the equation (5)

$$W_{wi}(t+1) = W_{wi}(t) + \alpha(t)(X_i - W_{wi}(t)) \quad \textbf{(5)}$$

In the equation (5), the coefficient 'α' is known as the Kohonen learning rate. The training process starts with an initial learning rate value '0.0' that is gradually decreased during training according to the equation (6)

$$\alpha(t) = \alpha_o (1 - \frac{t}{T}) \quad \textbf{(6)}$$

where

T is the maximum iteration number of the stage 'A' of the algorithm.

A typical initial value for the Kohonen learning rate is a value of '0.7'.

Step 8: The steps 1 to 6 are repeated until all training patterns have been processed once. For each training pattern 'p', the distance 'D_p' of the winning neuron is stored for further processing. The storage of this distance is performed before the weight update operation.

Step 9: At the end of each epoch, the training set mean error is calculated according to the equation (7)

$$E_i = \frac{1}{p} \sum_{k=1}^{p} D_k \quad \textbf{(7)}$$

where

P is the number of pairs in the training set,

D_k is the distance of the winning neuron for the pattern 'k' and 'i' is the current training epoch.

The network converges when the error measure falls below a user supplied tolerance value. The network also stops training in the case where the specified number of iterations has been performed, but the error value has not converged to a specific value.

Training of the weights from the hidden to the output nodes

Step 1: The synaptic weights of the network between the Kohonen and the Grossberg layer are set to small random values in the interval [0, 1].

Step 2: A vector pair (x, y) of the training set, is selected in random.

Step 3: The input vector 'x' of the selected training pattern is normalized

Step 4: The normalized input vector is sent to the network.

Step 5: In the hidden competitive layer the distance between the weight vector and the current input vector is calculated for each hidden neuron 'j' according to the equation (4).

Step 6: The winner neuron 'W' of the Kohonen layer is identified as the neuron with the minimum distance value 'D_j'. The output of this node is set to unity while the outputs of the other hidden nodes are assigned to zero values.

Step 7: The connection weights between the winning neuron of the hidden layer and all 'N' neurons of the output layer are adjusted according to the equation

$$V_{jw}(t+1) = V_{jw}(t) + \beta (y_j - V_{jw}(t)) \quad \textbf{(8)}$$

In the equation (8), the 'β' coefficient is known as the Grossberg learning rate.

Step 8: The above procedure is performed for each training pattern. In this case, the error measure is computed as the mean Euclidean distance between the winner node's output weights and the desired output, that is

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$E = \frac{1}{p}\sum_{j=1}^{N} D_j = \frac{1}{p}\sum_{j=1}^{P}\sum_{k=1}^{N}\sqrt{(y_k - w_{kj})^2} \quad (9)$$

IV. RESULTS AND DISCUSSION

Providing a machine the ability to see and understand as humans do has long fascinated scientists, engineers and even the common man. Synergistic research efforts in various scientific disciplines–Computer Vision, Artificial Intelligence, Neuroscience, Linguistics have brought us closer to this goal than at any other point in history. In our system we use four different subjects as they have lot of sequences to work with. Each HMM is created and trained using four walking sequences from the subject. Here, four persons walking sequences are taken where three person walking sequence belongs to original data. Each person walking sequence is of four files. So, it infers that out of sixteen files, eight files belong to original data and other eight files belong to other person's data. After testing, eight files will match with the original data and other eight are mismatched. Therefore, out of sixteen files eight are correct and eight are incorrect.

V. CONCLUSIONS

Most of the approaches for recognizing and detecting action and activities are based on the premise that the action /activity has already occurred. Reasoning about the intentions of humans and inferring what is going to happen presents a significant intellectual challenge. Security applications are among the first that stand to benefit from such a system, where detection of threat is of utmost importance. This paper has implemented GAIT recognition using HMM and CPN neural networks. The training patterns are generated from the frames of the walking sequence taken as video. HMM is used for extracting features from the frames and these features of each frame are used as training and testing patterns.

REFERENCES

[1] Aggarwal J., and Cai Q., 1999, Human motion analysis: a review, Computer Vision and Image Understanding, Vol.73, No.3, pp.428–440.

[2] Athanasios, I. Margaris, and Efthimios Kotsialos, 2004, Parallel counter propagation networks, In Proceedings of the International Conference on Theory and Applications of Mathematics and Informatics, Greece, pp.306-325.

[3] Bobick A., and Johnson A., 2001, Gait recognition using static activity-specific parameters. Proc. of IEEE Conference on Computer Vision and Pattern Recognition (Lihue, HI), December.

[4] Cunado D., Nash J., Nixon M., and Carter J.N., 1995, Gait extraction and description by evidence-gathering. Proc. of the International Conference on Audio and Video Based Biometric Person Authentication, pp.43–48.

[5] Gavrila D. M., 1999, The visual analysis of human movement: a survey, Computer Vision and Image Understanding, Vol. 73, No. 1, pp. 82–98.

[6] Huang P., Harris C., and Nixon M., 1999, Recognizing humans by gait via parametric canonical space, Artificial Intelligence in Engineering, Vol.13, No.4, pp.359–366.

[7] James A. Freeman, and David M. Skapura, 1991, Neural Networks: Algorithms, Applications, and Programming Techniques, Addison-Wesley Publishing Company, ISBN 0-201-51376-5.

[8] Johansson, G., 1973, Visual perception of biological motion and a model for its analysis, Perception and Psychophysics, Vol. 14, No.2, pp. 201–211. [3] Cedras, C., and Shah, M., 1995, Motion-based recognition: A survey, Image and Vision Computing, Vol.13, No.2, pp. 129–155.

[9] Little J., and Boyd J., 1998, Recognizing people by their gait: the shape of motion, Videre: Journal of Computer Vision Research, Winter 1998, Vol.1, No.2, pp.1–32.

[10] Milovanovic I., 2008, Radial Basis Function (RBF) networks for improved gait analysis, 9th Symposium

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

on Neural Network Applications in Electrical Engineering, 25-27 September, pp.19-132.

[11] Moeslund T. B., Hilton A., and Kruger V., 2006, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding, Vol. 104, No. 2, pp.90–126.

[12] Murase H., and Sakai R., 1996, Moving object recognition in eigenspace representation: gait analysis and lip reading, Pattern Recognition Letters, Vol.17, pp.155–162.

[13] Murray M., Drought A., and Kory R., 1964, Walking patterns of normal men, Journal of Bone and Joint surgery, Vol.46-A, No.2, pp.335–360.

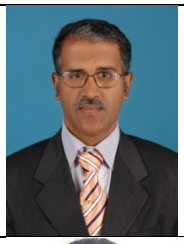
[14] Sasi varnan, C., Jagan, A., Jaspreet Kaur, Divya Jyoti and Rao, D.S., 2011, Gait Recognition Using Extracted Feature Vectors, IJCST, Vol. 2, Issue 3, pp. 77-84.

[15] Starner T., Weaver J., and Pentland A., 1998, Real-time American sign language recognition from video using HMMs, IEEE Trans. on Pattern Anal., and Machine Intell., Vol.12, No. 8, pp.1371–1375.

[16] Wilson D., and Bobick A., 1998, Non-linear pHMMs for the interpretation of parameterized gesture, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (Santa Barbara, CA), June.

[17] Yamato J., Ohya J., and Ishii L., 1995, Recognizing human action in time-sequential images using hidden Markov model, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.624–630.

P.Sripriya , Research scholar, Vels University has completed MCA from D G Vaishnav College, Chennai, India and completed M.Phil, from Algappa University, India. She has 12 years of teaching experience and published papers in National and International conferences. She is presently working as Asst. Professor in Department of MCA, School of Computing Sciences, VELS University, Chennai.



Dr.S.Purushothaman completed his PhD from Indian Institute of Technology Madras, India in 1995. He has 129 publications to his credit. He has 19 years of teaching experience. Presently he is working as Professor in PET Engineering College, India

R.Rajeswari completed M.Sc., Information Technology from Bharathidasan university, Tiruchirappalli and M.Phil Computer Science from Alagappa University, Karaikudi, Tamilnadu, India. She is currently pursuing PhD in Mother Teresa Women's University. Her area of interest is Intelligent Computing, image processing.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A Brief Study of Data Compression Algorithms

Yogesh *Rathore*
CSE,UIT, RGPV
Bhopal, M.P., India

Manish k. Ahirwar
CSE,UIT, RGPV
Bhopal, M.P., India

Rajeev Pandey
CSE,UIT, RGPV
Bhopal, M.P., India

***Abstract*—This paper present survey of several lossless data compression techniques and its corresponding algorithms. A set of selected algorithms are studied and examined. This paper concluded by stating which algorithm performs well for text data.**

Keywords-Compression; Encoding; REL; RLL; Huffman; LZ; LZW;

I. INTRODUCTION

In 1838 morse code used data compression for telegraphy which was based on using shorter code words for letters such as "e" and "t" that are more common in English . Modern work on data compression began in the late 1940 s with the development of information theory.

In 1949 Claude Shannon and Robert Fano devised a systematic way to assign code words based on probabilities of blocks. In 1951 David Huffmann found an optimal method for Data Compression. Early implementations were typically done in hardware, with distinct choices of code words being made as compromises between compression and error correction. With online storage of text file becoming general, software compression programs began to be developed IN EARLY 1970S , almost all COMPRESSIONS were based on adaptive Huffman coding. In the late 1980s, digital images became more generic, and standards for compressing them emerged, lossy compression methods also began to be widely used In the early 1990s. Current image compression standards include:FAX CCITT 3 (run-length encoding, with code words determined by Huffman coding from a definite distribution of run lengths); GIF (LZW); JPEG (lossy discrete cosine transform, then Huffman or arithmetic coding); BMP (run-length encoding, etc.); TIFF (FAX, JPEG, GIF, etc.).With the growing demand for text transmission and storage due to advantage of Internet technology, text compression has become most important part of computer technology. Compression is used to solve this problem by reducing the file size without affecting the quality of the original Data.

With this trend expected to run, it makes sense to pursue research on developing algorithms that can most effectively use available network bandwidth with maximally compressing data. It is also necessary to consider the security aspects of the data being transmitted while compressing it, as most of the text information transmitted over the Internet is very much vulnerable to a mass of attacks. Researchers have developed highly sophisticated approaches for lossless text compression , such as Huffman encoding, arithmetic encoding, the Lempel-Ziv etc.

Compression methods have a long list. In this paper, we shall discuss only the lossless text compression techniques and not the lossy techniques as related to our work. In this, reviews of different basic lossless text data compression methods are considered. The methods such as Run Length Encoding, Huffman coding, Shannon-Fano Coding and Arithmetic coding are considered. Lempel Ziv scheme is also considered which a dictionary based technique. A conclusion is derived on the basis of these methods based software.

II. COMPRESSION & DECOMPRESSION

Compression is a technology by which one or more files or directory size can be reduced so that it is easy to handle. The objective of compression is to reduce the number of bits required to represent data and to decrease the transmission time. Compression is achieved through encoding data and the data is decompressed to its original form by decoding. Compression increases the capacity of a communication channel by transmitting the compressed file. A common compressed file which is used day-today has extensions which end with .Sit, .Tar, .Zip;

There are two main types of data compression: lossy and lossless.

A. Lossless Compression Techniques

Lossless compression techniques resurface the original data from the compressed file without any loss of data. Thus the information does not alter during the compression and decompression processes. Lossless compression techniques are used to compress images, text and medical images preserved for juristic reasons, computer executable file and so on.

B. Lossy compression techniques

Lossy compression techniques resurface the original message with loss of some information. It is not possible to resurface the original message using the decoding process. The decompression process results an nearly realignment. It may be desirable, when data of some ranges which could not recognized by the human brain can be ignored. Such techniques could be used for multimedia audio, video and images to achieve more compact data compression.

Compression is a technology by which one or more files or directory size can be reduced so that it is easy to handle. The objective of compression is to reduce the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

number of bits required to represent data and to decrease the transmission time. Compression is achieved through encoding data and the data is decompressed to its original form by decoding. Compression increases the capacity of a communication channel by transmitting the compressed file. A common compressed file which is used day-today has extensions which end with .Sit, .Tar, .Zip;

III. COMPRESSION TECHNIQUES

Many different techniques are used to compress data. Most compression techniques cannot stand on their own, but must be combined together to form a compression algorithm. Those that can stand alone are often more effective when joined together with other compression techniques. Most of these techniques fall under the category of entropy coders, but there are others such as Run-Length Encoding and the Burrows-Wheeler Transform that are also commonly used. Compression techniques have a long list. In this paper, we shall discuss only the lossless compression techniques and not the lossy techniques as related to our work.

A. Run Length Encoding Algorithm

Run Length Encoding or simply RLE is the simplest of the data compression algorithms. The consecutive sequences of symbols are identified as runs and the others are identified as non runs in this algorithm. This algorithm deals with some sort of redundancy. [14] It checks whether there are any repeating emblem or not, and is based on those redundancies and their length. Continuously recurrent symbols are identified as runs and all the other sequences are considered as non-runs. For an example, the text "ABABBBBC" is considered as a source to compress, then the first three letters are considered as a non-run with length three, and the next 4 letters are considered as a run with length 4 since there is a repetition of symbol B.

The major task of this algorithm is to identify the runs of the source file and to record the symbol and the length of each run. The Run Length Encoding algorithm uses those runs to compress the original source file while keeping all the non-runs with out using for the compression process. [14]

B. Huffman Encoding

Huffman Encoding Algorithms use the probability distribution of the alphabet of the source to develop the code words for symbols. The repetition distribution of all the characters of the source is calculated in order to calculate the probability distribution. The code words are assigned pursuant to the probabilities. Smaller code words for higher probabilities and longer code words for smaller probabilities are assigned. For this work a binary tree is created using the symbols as leaves according to their probabilities and paths of those are taken as the code words.

The two approaches of Huffman Encoding have been proposed first is Static Huffman Algorithms and the second one is Adaptive Huffman Algorithms.

Static Huffman Algorithms compute the frequencies first and then generate a common tree for both the compression and decompression processes . Details of this tree should be saved or transferred to the compressed file.

The Adaptive Huffman algorithms develop the tree while calculating the frequencies and there will be two trees in both the processes. In this method, a tree is generated with the flag symbol in the beginning and is updated as the next symbol is read.

C. The Lempel Zev Welch Algorithm

Dictionary based compression algorithms are based on a dictionary instead of a statistical model .

LZW is the most popular method. This technique has been applied for data compression.

The main steps for this technique are given below:-

1. Firstly it will read the file and given a code to each character.

2. If the same characters are found in a file then it will not assign the new code and then use the existing code from a dictionary.

3. The process is continuous until the characters in a file are null..

The application software that makes the use of Lampel Zev Welch algorithm is "LZIP". Which makes the use of the dictionary based compression method.

D. Burrows-Wheeler Transform

The Burrows-Wheeler Transform is a compression technique invented in 1994 that aims to reversibly transform a block of input data such that the amount of runs of identical characters is maximized. The BWT itself does not perform any compression operations, it simply transforms the input such that it can be more efficiently coded by a Run-Length Encoder or other secondary compression technique.

The algorithm for a BWT is as follows:
1. Create a string array.
2. Generate all produce rotations of the input string, storing every within the array.
3. Kind the array alphabetically.
4. Come the last column of the array.

BWT usually works best on long inputs with many alternating identical characters. Here is an example of the algorithm being run on an ideal input.

TABLE I. EXAMPLE OF BURROWS-WHEELER TRANSFORM

Input	Rotations	Alpha-Sorted Rotations	Output
HAHAHA&	HAHAHA&	AHAHA&**H**	HHH&AAA
	&HAHAHA	AHA&HA**H**	
	A&HAHAH	A&HAHA**H**	
	HA&HAHA	HAHAHA**&**	
	AHA&HAH	HAHA&H**A**	
	HAHA&HA	HA&HAH**A**	
	AHAHA&H	&HAHAH**A**	

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Because of its alternating identical characters, performing the BWT on this input generates an optimal result that another algorithm could further compress, such as RLE which would yield "3H&3A". While this example generated an optimal result, it does not generate optimal results on most real-world data.

E. Shannon-Fano Coding

This is one of the earliest compression techniques, invented in 1949 by Claude Shannon and Robert Fano. This technique involves generating a binary tree to represent the probabilities of each symbol occurring. The symbols are ordered such that the most frequent symbols appear at the top of the tree and the least likely symbols appear at the bottom.

The code for a given symbol is obtained by searching for it in the Shannon-Fano tree, and appending to the code a value of 0 or 1 for each left or right branch taken, respectively. For example, if “A” is two branches to the left and one to the right its code would be “0012”. Shannon-Fano coding does not always produce optimal codes due to the way it builds the binary tree from the bottom up. For this reason, Huffman coding is used instead as it generates an optimal code for any given input.

The algorithm to generate Shannon-Fano codes is fairly simple

[1] Parse the input, counting the occurrence of each symbol.

[2] Determine the probability of each symbol using the symbol count.

[3] Sort the symbols by probability, with the most probable first.

[4] Generate leaf nodes for each symbol.

[5] Divide the list in two while keeping the probability of the left branch roughly equal to those on the right branch.

[6] Prepend 0 and 1 to the left and right nodes' codes, respectively.

[7] Recursively apply steps 5 and 6 to the left and right subtrees until each node is a leaf in the tree. [15]

F. Arithmetic Coding

This method was developed in 1979 at IBM, which was investigating data compression techniques for use in their mainframes. Arithmetic coding is arguably the most optimal entropy coding technique if the objective is the best compression ratio since it usually achieves better results than Huffman Coding. It is, however, quite complicated compared to the other coding techniques.

Rather than splitting the probabilities of symbols into a tree, arithmetic coding transforms the input data into a single rational number between 0 and 1 by changing the base and assigning a single value to each unique symbol from 0 up to the base. Then, it is further transformed into a fixed-point binary number which is the encoded result. The value can be decoded into the original output by changing the base from binary back to the original base and replacing the values with the symbols they correspond to.

A general algorithm to compute the arithmetic code is:

- Calculate the number of unique symbols in the input. This number represents the base b (e.g. Base 2 is binary) of the arithmetic code.
- Assign values from 0 to b to each unique symbol in the order they appear.
- Using the values from step 2, replace the symbols in the input with their codes
- Convert the result from step 3 from base b to a sufficiently long fixed-point binary number to preserve precision.
- Record the length of the input string somewhere in the result as it is needed for decoding.

Here is an example of an encode operation, given the input “ABCDAABD”:

[1] Found 4 unique symbols in input, therefore base = 4. Length = 8

[2] Assigned values to symbols: A=0, B=1, C=2, D=3

[3] Replaced input with codes: “0.012300134” where the leading 0 is not a symbol.

[4] Convert “0.012311234” from base 4 to base 2: “0.011011000001112”

[5] Result found. Note in result that input length is 8.

Assuming 8-bit characters, the input is 64 bits long, while its arithmetic coding is just 15 bits long resulting in an excellent compression ratio of 24%. This example demonstrates how arithmetic coding compresses well when given a limited character set.

IV. COMPRESSION ALGORITHAMS

Many different techniques are used to compress data. Most compression techniques cannot stand on their own, but must be combined together to form a compression algorithm. These compression algorithms are described as follows:

A. Sliding Window Algorithms

1) LZ77

Published in 1977, LZ77 is the algorithm that started it all. It introduced the concept of a 'sliding window' for the first time which brought about significant improvements in compression ratio over more primitive algorithms.

LZ77 maintains a dictionary using triples representing offset, run length, and a deviating character. The offset is how far from the start of the file a given phrase starts at, and the run length is how many characters past the offset are part of the phrase.

The deviating character is just an indication that a new phrase was found, and that phrase is equal to the phrase from

Identify applicable sponsor/s here. *(sponsors)*

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

offset to offset+length plus the deviating character. The dictionary used changes dynamically based on the sliding window as the file is parsed. For example, the sliding window could be 64MB which means that the dictionary will contain entries for the past 64MB of the input data.

Given an input "abbadabba" the output would look something like "abb(0,1,'d')(0,3,'a')" as in the example below:

TABLE II. EXAMPLE OF LZ77

Position	Symbol	Output
0	a	A
1	b	b
2	b	b
3	a	(0, 1, 'd')
4	d	
5	a	(0, 3, 'a')
6	b	
7	b	
8	a	

While this substitution is slightly larger than the input, it usually achieves a significantly smaller result given longer input data. [3]

2) LZR

LZR is a modification of LZ77 invented by Michael Rodeh in 1981. The algorithm aims to be a linear time alternative to LZ77. However, the encoded pointers can indicate to any offset in the file which means LZR consumes a considerable amount of memory. Together with its poor compression ratio (LZ77 is often superior) it is an unfeasible variant.[18][19]

3) DEFLATE

DEFLATE was invented by Phil Katz in 1993 and is the basis for the majority of compression tasks today. It simply combines an LZ77 or LZSS preprocessor with Huffman coding on the back end to achieve moderately compressed results in a short time.

It is used in gzip software. It is use .gz extension Its compression quantitative relation show area unit show below

File : Example1.doc
File Size : 7.0 MB
Compressed File Size : 1.8 MB

File : Example2.doc
File Size : 1.1 MB
Compressed File Size : 854.0 KB

File : Example3.pdf
File Size : 453 KB
Compressed File Size : 369.7 KB

File : Example4.txt
File Size : 71.1 KB
Compressed File Size : 14.8 KB

File : Example5.doc
File Size : 599.6 MB
Compressed File Size : 440.6 KB

4) DEFLATE64

DEFLATE64 is a proprietary extension of the DEFLATE algorithm which increases the dictionary size to 64kB (hence the name) and allows greater distance in the sliding window. It increases both performance and compression ratio compared to DEFLATE. [20] However, the proprietary nature of DEFLATE64 and its modest improvements over DEFLATE has led to limited adoption of the format. Open source algorithms such as LZMA are generally used instead.

5) LZSS

The LZSS, or Lempel-Ziv-Storer-Szymanski algorithm was First published in 1982 by James Storer and Thomas Szymanski. LZSS ameliorate on LZ77 in that it can detect whether a substitution will decrease the file size or not.

If no size reduction is going to be achieved, the input is left as a literal within the output. Otherwise, the section of the input is replaced with an (offset, length) pair where the offset is how many bytes from the start of the input and the length is how many characters to read from that position. [21] Another improvement over LZ77 comes from the elimination of the "next character" and uses just an offset-length pair.

Here is a brief example given the input " these theses" which yields " these (0,6) s" which saves just one byte, but saves considerably more on larger inputs.

TABLE III. EXAMPLE OF LZSS

Index	0	1	2	3	4	5	6	7	8	9	10	11	12
Symbol		t	h	e	s	e		t	h	e	s	e	s
Substituted		t	h	e	s	e	(	0	,	6	)	s	

LZSS is still used in many popular archive formats, the best known of which is RAR. LZSS is also sometimes used for network data compression.

6) LZH

LZH was developed in 1987 and it stands for "Lempel-Ziv Huffman." It is a variant of LZSS that utilizes Huffman coding to compress the pointers, resulting in inchmeal better compression. However, the improvements gained using Huffman coding are negligible and the compression is not worth the performance hit of using Huffman codes. [19]

7) LZB

LZB was also developed in 1987 by Timothy Bell et al as a variant of LZSS. Like LZH, LZB also aims to reduce the compressed file size by encoding the LZSS pointers more efficiently. The way it does this is by gradually increasing

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

the size of the pointers as the sliding window grows larger. It can achieve higher compression than LZSS and LZH, but it is still rather slow as compared to LZSS due to the extra encoding step for the pointers. [19]

8) ROLZ

ROLZ stands for "Reduced Offset Lempel-Ziv" and its goal is to improve LZ77 compression by restricting the offset length to reduce the amount of data required to encode the offset-length pair. This derivative of LZ77 was first seen in 1991 in Ross Williams' LZRW4 algorithm. Other implementations include BALZ, QUAD, and RZM. Highly optimized ROLZ can achieve nearly the same compression ratios as LZMA; however, ROLZ suffers from a lack of popularity.

9) LZP

LZP stands for "Lempel-Ziv combined with Prediction." It is a special case of ROLZ algorithm where the offset is reduced to 1. [14] There are several variations using different techniques to achieve either faster operation of better compression ratios. LZW4 implements an arithmetic encoder to achieve the best compression ratio at the cost of speed. [22]

10) LZRW1

Ron Williams created this algorithm in 1991, introducing the concept of a Reduced-Offset Lempel-Ziv compression for the first time. LZRW1 can achieve high compression ratios while remaining very fast and efficient. Ron Williams also created several variants that improve on LZRW1 such as LZRW1-A, 2, 3, 3-A, and 4. [23]

11) LZJB

Jeff Bonwick created his Lempel-Ziv Jeff Bonwick algorithm in 1998 for use in the Solaris Z File System (ZFS). It is considered a variant of the LZRW algorithm, specifically the LZRW1 variant which is aimed at maximum compression speed. Since it is used in a file system, speed is especially important to ensure that disk operations are not bottlenecked by the compression algorithm.

12) LZS

The Lempel-Ziv-Stac algorithm was developed by Stac Electronics in 1994 for use in disk compression software. It is a modification to LZ77 which distinguishes between literal symbols in the output and offset-length pairs, in addition to removing the next encountered symbol. The LZS algorithm is functionally most similar to the LZSS algorithm. [24]

13) LZX

The LZX algorithm was developed in 1995 by Jonathan Forbes and Tomi Poutanen for the Amiga computer. The X in LZX has no Special meaning. Forbes sold the algorithm to Microsoft in 1996 and went to work for them, where it was further improved upon for use in Microsoft's cabinet (. CAB) format. This algorithm is also employed by Microsoft to compress Compressed HTML Help (CHM) files, Windows Imaging Format (WIM) files, and Xbox Live Avatars. [25]

14) LZO

LZO was developed by Markus Oberhumer in 1996 whose development goal was fast compression and decompression. It allows for adjustable compression levels and requires only 64kB of additional memory for the highest compression level, while decompression only requires the input and output buffers. LZO functions very similarly to the LZSS algorithm but is optimized for speed rather than compression ratio.

15) LZMA

The Lempel-Ziv Markov chain Algorithm was published in 1998 with the release of the 7-Zip archiver for use in the .7z file format. It achieves better compression than bzip2, DEFLATE, and other algorithms in most cases. LZMA uses a chain of compression techniques to achieve its output. First, a modified LZ77 algorithm, which operates at a bitwise level rather than the traditional bytewise level, is used to parse the data. Then, the output of the LZ77 algorithm undergoes arithmetic coding. More techniques can be applied depending on the specific LZMA implementation. The result is considerably improved compression ratios over most other LZ variants mainly due to the bitwise method of compression rather than bytewise.[27]

It is used in 7zip software. It uses .7z extension. Its compression quantitative relation show area unit shows below

File	: Example1.doc
File Size	: 7.0 MB
Compressed File Size	: 1.2 MB
File	: Example2.doc
File Size	: 1.1 MB
Compressed File Size	: 812.3 KB
File	: Example3.pdf
File Size	: 453 KB
Compressed File Size	: 365.7 KB
File	: Example4.txt
File Size	: 71.1 KB
Compressed File Size	: 12.4 KB
File	: Example5.doc
File Size	: 599.6 MB
Compressed File Size	: 433.5 KB

a) LZMA2

LZMA2 is an incremental improvement to the original LZMA algorithm, first introduced in 2009 [28] in an update to the 7-Zip archive software. LZMA2 improves the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

multithreading capabilities and thus the performance of the LZMA algorithm, as well as better handling of incompressible data resulting in slightly better compression.

It is used in Xzip software. It is use .xz extension. Its compression quantitative relation show area unit shows below

File : Example1.doc
File Size : 7.0 MB
Compressed File Size : 1.2 MB

File : Example2.doc
File Size : 1.1 MB
Compressed File Size : 811.9 KB

File : Example3.pdf
File Size : 453 KB
Compressed File Size : 365.7 KB

File : Example4.txt
File Size : 71.1 KB
Compressed File Size : 12.4 KB

File : Example5.doc
File Size : 599.6 MB
Compressed File Size : 431.0 KB

b) Statistical Lempel-Ziv

Statistical Lempel-Ziv was a concept created by Dr. Sam Kwong and Yu Fan Ho in 2001. The basic principle it operates on is that a statistical analysis of the data can be combined with an LZ77-variant algorithm to further optimize what codes are stored in the dictionary.

It is used in LZMA software. It uses .lzma extension. Its compression quantitative relation show area unit shows below

File : Example1.doc
File Size : 7.0 MB
Compressed File Size : 1.2 MB

File : Example2.doc
File Size : 1.1 MB
Compressed File Size : 812.1 KB

File : Example3.pdf
File Size : 453 KB
Compressed File Size : 365.6 KB

File : Example4.txt
File Size : 71.1 KB
Compressed File Size : 12.3 KB

File : Example5.doc
File Size : 599.6 MB
Compressed File Size : 433.3 KB

B. Dictionary Algorithms

1) LZ78

LZ78 was created by Lempel and Ziv in 1978. Rather than using a sliding window to generate the dictionary, the input information is either preprocessed to generate a dictionary with the infinite scope of the input, or the dictionary is built up as the file is parsed. LZ78 employs the latter strategy. The dictionary size is usually limited to a few MB, or all codes up to a certain number of bytes such as 8; this is done to reduce memory requirements. How the algorithm controls the dictionary being full is what sets most LZ78 type algorithms apart. [4]

While parsing the file, the LZ78 algorithm adds each newly encountered a character or string of characters to the dictionary. For each symbol in the input, a dictionary entry in the form (dictionary index, unknown symbol) is generated; if a symbol is already in the dictionary then the dictionary will be searched for substrings of the current symbol and the symbols following it.

The index of the longest substring match is used for the dictionary index. The information show to by the dictionary index is added to the final character of the obscure substring. if the present image is obscure, then the concordance file is situated to 0 to demonstrate that it is a solitary character passage. The section's structure a connected record sort information structure.

An input such as "xyyxyxyyxxyxxy" would generate the output {(0,x)(0,y)(2,x)(0,y)(1,y)(3,x)(6,y)}. You can see how this was derived in the following example:

TABLE IV. EXAMPLE OF LZ78

Input:		x	b	bx	d	xb	bxx	bxxd
Dictionary Index	0	1	2	3	4	5	6	7
Output	NULL	(0,x)	(0,b)	(2,x)	(0,d)	(1,b)	(3,x)	(6,d)

2) LZW

LZW is the Lempel-Ziv-Welch algorithm created in 1984 by Terry Welch. It is the most commonly used derivative of the LZ78 family, nevertheless being heavily patent-encumbered. LZW improves on LZ78 in a similar way to

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

LZSS; it removes redundant characters in the output and makes the output entirely out of pointers. It also includes every character in the dictionary before starting compression, and employs other tricks to improve compression such as encoding the last character of every new phrase as the first character of the next phrase. LZW is commonly found in the GIF Format, as well as in the early specificiations of the ZIP format and other specialized applications.

LZW is very fast, but achieves low compression compared to most newer algorithms and some algorithms are both faster and achieve better compression.

It is used in WinZip software. It is use .zip extension. Its compression quantitative relation show area unit shows below

File	: Example1.doc
File Size	: 7.0 MB
Compressed File Size	: 1.7 MB

File	: Example2.doc
File Size	: 1.1 MB
Compressed File Size	: 851.6 KB

File	: Example3.pdf
File Size	: 453 KB
Compressed File Size	: 368.4 KB

File	: Example4.txt
File Size	: 71.1 KB
Compressed File Size	: 14.2 KB

File	: Example5.doc
File Size	: 599.6 MB
Compressed File Size	: 437.3 KB

3) LZC

LZC, or Lempel-Ziv Compress is a slight modification to the LZW algorithm used in the UNIX compress utility.

The main difference between LZC and LZW is that LZC monitors the compression ratio of the output. Once the ratio crosses a certain threshold, the dictionary is rejected and rebuilt. [19]

4) LZAP

LZAP was created in 1988 by James Storer as a modification to the LZMW algorithm. The AP stands for "all prefixes" in that rather than storing a single phrase in the dictionary each iteration, the dictionary stores every permutation. For example, if the last phrase was "last" and the current phrase is "next" the dictionary would store "lastn", "lastne", "lastnex", and "lastnext".

5) LZWL

LZWL is a revisal to the LZW algorithm created in 2006.It works with syllables rather than a character. LZWL is designed to work better with certain data sets with many commonly occurring syllables such as XML data. That algorithm is usually used with a preprocessor that decomposes the input data into syllables. [31]

6) LZJ

Matti Jakobsson published the LZJ algorithm in 1985 [32] and it is one of the only LZ78 algorithms that deviates from LZW. The methods works by storing every unique string in the already processed input up to an arbitrary maximum length in the dictionary and assigning codes to every. When the dictionary is full, all entries that occurred only once are removed. [19]

C. Non-dictionary Algorithms

1) PPM

Prediction by Partial Matching is a statistical modeling technique that uses a set of previous symbols in the input to predict what the next symbol will be in order to reduce the entropy of the output data. This is different from a dictionary since PPM makes predictions about what the next symbol will be rather than trying to find the next symbols in the dictionary to code them. PPM is usually combined with an encoder on the back end, such as arithmetic coding or adaptive Huffman coding. [33] PPM or a variant known as PPM are implemented in many archive formats including 7-Zip and RAR.

It is used in RAR software. It uses .rar extension. Its compression quantitative relation show area unit shows below

File	: Example1.doc
File Size	: 7.0 MB
Compressed File Size	: 1.4 MB

File	: Example2.doc
File Size	: 1.1 MB
Compressed File Size	: 814.5 KB

File	: Example3.pdf
File Size	: 453 KB
Compressed File Size	: 367.7 KB

File	: Example4.txt
File Size	: 71.1 KB
Compressed File Size	: 13.9 KB

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

File : Example5.doc

File Size : 599.6 MB

Compressed File Size : 436.9 KB

2) bzip2

bzip2 is an open source implementation of the Burrows-Wheeler Transform. Its operating principles are simple, yet they achieve a very good compromise between speed and compression ratio that makes the bzip2 format very popular in UNIX environments. First, a Run-Length Encoder is applied to the data. Next, the Burrows-Wheeler Transform is applied. Then, a move-to-front transform is applied with the intent of creating a large amount of identical symbols forming runs for use in yet another Run-Length Encoder. Finally, the result is Huffman coded and wrapped with a header. [34]

It is used in bzip2 software. It uses .bz2 extension. Its compression quantitative relation show area unit shows below

File : Example1.doc

File Size : 7.0 MB

Compressed File Size : 1.5 MB

File : Example2.doc

File Size : 1.1 MB

Compressed File Size : 871.8 KB

File : Example3.pdf

File Size : 453 KB

Compressed File Size : 374.1 KB

File : Example4.txt

File Size : 71.1 KB

Compressed File Size : 15.5 KB

File : Example5.doc

File Size : 599.6 MB

Compressed File Size : 455.9 KB

3) PAQ

PAQ was created by Matt Mahoney in 2002 as an improvement upon older PPM(d) algorithms. The way it does this is by using a revolutionary technique called context mixing. Context mixing is when multiple statistical models (PPM is one example) are intelligently combined to make better predictions of the next symbol than either model by itself. PAQ is one of the most promising algorithms because of its extremely high compression ratio and very active development. Over 20 variants have been created since its inception, with some variants achieving record compression ratios. The biggest drawback of PAQ is its slow speed due to using multiple statistical models to get the best compression ratio. However, since hardware is constantly getting faster, it may be the standard of the future. [35] PAQ is slowly being adopted, and a different called PAQ8O which brings 64-bit support and major speed improvements can be found in the PeaZip program for Windows. Other PAQ formats are mostly command-line only.

V. CONCLUSION

There we talked about a need of data compression, and situations in which these lossless methods are useful. The algorithms used for lossless compression are described in brief. A special, Run-length coding, statistical encoding and dictionary based algorithm like LZW, are provided to the concerns of this family compression method. In the Statistical compression techniques, Arithmetic coding technique performs with an improvement over Huffman coding, over Shannon-Fano coding and over Run Length Encoding technique. Compression techniques improve the efficiency compression on text data. Lempel-Ziv Algorithm is best of these Algorithms.

TABLE V. COMPRESSION

Compression Software Extension	**Example 1.doc (7.0 MB)**	**Example 2.doc (1.1 MB)**	**Example 3.pdf (453 KB)**	**Example 4 .txt (71.7 KB)**	**Example 5.doc (599.6 KB)**
	After Compression File Size				
.7z	1.2 MB	812.3 kB	365.7 kB	12.4 kB	433.5 kB
.bz2	1.5 MB	871.8 kB	374.1 kB	15.5 kB	455.9 kB
.gz	1.8 MB	854.0 kB	369.7 kB	14.8 kB	440.6 kB
.lzma	1.2 MB	812.1 kB	365.6 kB	12.3 kB	433.3 kB
.xz	1.2 MB	811.9 kB	365.7 kB	12.4 kB	431.0 kB
.zip	1.7 MB	851.6 kB	368.4 kB	14.2 kB	437.3 kB
.rar	1.4 MB	814.5 kB	367.7 kB	13.9 kB	436.9 kB

REFERENCES

[1] Lynch, Thomas J., Data Compression: Techniques and Applications,Lifetime Learning Publications, Belmont, CA, 1985

[2] Philip M Long., Text compression via alphabet representation

[3] Cappellini, V., Ed. 1985. Data Compression and Error Control

Techniques with Applications. Academic Press, London.

[4] Cortesi, D. 1982. An Effective Text-Compression Algorithm. BYTE 7,1 (Jan.), 397-403.

[5] Glassey, C. R., and Karp, R. M. 1976. On the Optimality of Huffman Trees. SIAM J. Appl. Math 31, 2 (Sept.), 368-378.

[6] Knuth, D. E. 1985. Dynamic Huffman Coding. J. Algorithms 6, 2 (June), 163-180.

[7] Llewellyn, J. A. 1987. Data Compression for a Source with Markov Characteristics. Computer J. 30, 2, 149-156.

[8] Pasco, R. 1976. Source Coding Algorithms for Fast Data Compression.Ph. D. Dissertation, Dept. of Electrical Engineering, Stanford Univ., Stanford, Calif.

[9] Rissanen, J. J. 1983. A Universal Data Compression System. IEEE

Trans. Inform. Theory 29, 5 (Sept.), 656-664.

[10] Tanaka, H. 1987. Data Structure of Huffman Codes and Its Application to Efficient Encoding and Decoding. IEEE Trans. Inform. Theory 33,1 (Jan.), 154-156.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[11] Ziv, J., and Lempel, A. 1977. A Universal Algorithm for Sequential Data Compression. IEEE Trans. Inform. Theory 23, 3 (May), 337-343.

[12] Giancarlo, R., D. Scaturro, and F. Utro. 2009. Textual data compression in computational biology: a synopsis. Bioinformatics 25 (13): 1575-1586.

[13] Burrows M., and Wheeler, D. J. 1994. A Block-Sorting Lossless Data Compression Algorithm. SRC Research Report 124, Digital Systems Research Center.

[14] S. R. Kodifuwakku and U. S. Amarasinge, "Comparison of loosless data compression algorithms for text data".IJCSE Vol 1 No 4416-225.

[15] Shannon, C.E. (July 1948). "A Mathematical Theory of Communication". Bell System Technical Journal 27: 379–423. [16] HUFFMAN, D. A. 1952. A method for the construction of minimum-redundancy codes. In Proceedings of the Institute of Electrical and Radio Engineers 40, 9 (Sept.), pp. 1098-1101.

[17] RISSANEN, J., AND LANGDON, G. G. 1979. Arithmetic coding. IBM J. Res. Dev. 23, 2 (Mar.), 149-162.

[18] RODEH, M., PRATT, V. R., AND EVEN, S. 1981. Linear algorithm for data compression via string matching. J. ACM 28, 1 (Jan.), 16-24.

[19] Bell, T., Witten, I., Cleary, J., "Modeling for Text Compression", ACM Computing Surveys, Vol. 21, No. 4 (1989).

[20] DEFLATE64 benchmarks

[21] STORER, J. A., AND SZYMANSKI, T. G. 1982. Data compression via textual substitution. J. ACM 29, 4 (Oct.), 928-951.

[22] Bloom, C., "LZP: a new data compression algorithm", Data Compression Conference, 1996. DCC '96. Proceedings, p. 425 10.1109/DCC.1996.488353.

[23] http://www.ross.net/compression/

[24] "Data Compression Method - Adaptive Coding witih Sliding Window for Information Interchange", American National Standard for Information Systems, August 30, 1994.

[25] LZX Sold to Microsoft

[26] LZO Info

[27] LZMA Accessed on 12/10/2011.

[28] LZMA2 Release Date

[29] Kwong, S., Ho, Y.F., "A Statistical Lempel-Ziv Compression Algorithm for Personal Digital Assistant (PDA)", IEEE Transactions on Consumer Electronics, Vol. 47, No. 1, February 2001, pp 154-162.

[30] David Salomon, Data Compression – The complete reference, 4th ed., page 212

[31] Chernik, K., Lansky, J., Galambos, L., "Syllable-based Compression for XML Documents", Dateso 2006, pp 21-31, ISBN 80-248-1025-5.

[32] Jakobsson, M., "Compression of Character Strings by an Adaptive Dictionary", BIT Computer Science and Numerical Mathematics, Vol. 25 No. 4 (1985). doi>10.1007/BF01936138

[33] Cleary, J., Witten, I., "Data Compression Using Adaptive Coding and Partial String Matching", IEEE Transactions on Communications, Vol. COM-32, No. 4, April 1984, pp 396-402.

[34] Seward, J., "bzip2 and libbzip2", bzip2 Manual, March 2000.

[35] Mahoney, M., "Adaptive Weighting of Context Models for Lossless Data Compression", Unknown, 2002.

fMRI image Segmentation using conventional methods versus Contextual Clustering

[1]Suganthi D., and [2]Purushothaman S.,

[1]Suganthi D., Research Scholar,
Department of Computer Science,
Mother Teresa Women's University,
Kodaikanal, Tamilnadu, India-624101,

[2]Dr.Purushothaman S,
Professor, PET Engineering College, Vallioor, India-627117.

***Abstract*- Image segmentation plays a vital role in medical imaging applications. Many image segmentation methods have been proposed for the process of successive image analysis tasks in the last decades. The paper has considered fMRI segmentation inspite of existing techniques to segment the fMRI slices. In this paper an fmri image segmentation using contextual clustering method is presented. Matlab software 'regionprops' function has been used as one of the criteria to show performance of CC. The CC segmentaion shows more segmented objects with least discontuinty within the objects in the fMRI image. From the experimental results, it has been found that, the Contextual clustering method shows a better segmentation when compared to other conventional segmentation methods.**

Keywords*: *Contextual clustering; segmentation; fMRI image.

I. INTRODUCTION

Image segmentation is the process of partitioning / subdividing a digital image into multiple meaningful regions or sets of pixels regions with respect to a particular application. The segmentation is based on measurements taken from the image and might be grey level, color, texture, depth or motion. The result of image segmentation is a set of segments that collectively cover the entire image. All the pixels in region are similar with respect to some characteristic or computed property, such as color, intensity, or texture. For any object in an image, there are many 'features' which are interesting points on the object that can be extracted to provide a "feature" description of the object. Image segmentation is done using various edge detection techniques such as Sobel, Prewitt, Roberts, Canny, LoG,

MRI Segmentation provides great importance in research and clinical applications. There are many methods that exist to segment the brain. Conventional methods like sobel, canny, log, zerocross, and prewitt use pure image processing techniques that need human interaction for accurate and reliable segmentation. Unsupervised methods, can segment the brain with high precision. For this reason, unsupervised methods are preferred over conventional methods. Many unsupervised methods such as Fuzzy c-means, Finite Gaussian Mixture Model, Artificial Neural Networks, etc. are available.

II. RELATED WORK

Bueno et al. 2000, described an image-based method founded on mathematical morphology to facilitate the segmentation of cerebral structures on 3D magnetic resonance images (MRIs). Jose et al, 2003, descried parametric image segmentation that consists of finding a label field which defines a partition of an image into a set of non overlapping regions and the parameters of the models that describe the variation of some property within each region. A Bayesian formulation is presented, based on the key idea of using a doubly stochastic prior model for the label field, which allows one to find exact optimal estimators for both this field and the model parameters by the minimization of a differentiable function.

Liu, 2006, presented a new level set based solution for automatic medical image segmentation. Wee et al, 2006, described accurate segmentation of magnetic resonance (MR) images of the brain. They broadly divided current MR brain image segmentation algorithms into three categories: classification based, region-based, and contour-based. They showed that by incorporating two key ideas into the conventional fuzzy C-means clustering algorithm, they are able to take into account the local spatial context and compensate for the intensity non uniformity (INU) artifact during the clustering process. Xiangrong et al, 2010, described clustering

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

algorithms in tissue segmentation in MRI. The authors proposed an approach to tissue segmentation of 3D brain MRI using semi-supervised spectral clustering. Yan Li and Zheru Chi, 2005, described magnetic resonance imaging (MRI) as an advanced medical imaging technique providing rich information about the human soft tissue anatomy.

The goal of magnetic resonance (MR) image segmentation is to accurately identify the principal tissue structures in these image volumes. A new unsupervised MR image segmentation method based on self-organizing feature map (SOFM) network has been presented. Yongyue et al, 2001, stated that the finite mixture (FM) model is the most commonly used model for statistical segmentation of brain magnetic resonance (MR) images because of its simple mathematical form and the piecewise constant nature of ideal brain MR images. They proposed a hidden Markov random field (HMRF) model, which is a stochastic process generated by a MRF whose state sequence cannot be observed directly but which can be indirectly estimated through observations.

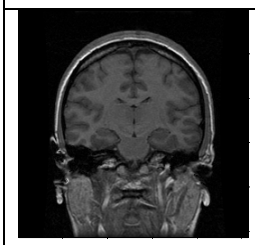
Zavaljevski et al, 2000 described MR brain image segmentation into several tissue classes is of significant interest to visualize and quantify individual anatomical structures. Zhang, 2004 stated that image segmentation plays a crucial role in many medical imaging applications. They presented a novel algorithm for fuzzy segmentation of magnetic resonance imaging (MRI) data

III. MATERIALS AND METHODOLOGY

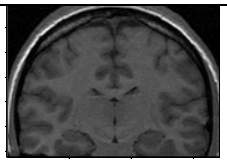
The Internet Brain Segmentation Repository (IBSR) provides manually-guided expert segmentation results along with magnetic resonance brain image data. fMRI slice images have been obtained from IBSR for use in this research work.

Table 1 presents two figures (full slice and cropped slice used for segmentation analysis).

Table 1 fMRI slice



Full slice image



Region cropped used for segmentation analysis

A Sobel Operator: It performs 2-D spatial gradient measurement on an image and so emphasizes regions of high spatial frequency that correspond to edges. The convolution masks of Sobel operator are as shown Table 2, which are used to obtain the gradient magnitude of the image from the original.

Table 2. Sobel Mask

1	2	1
0	0	0
-1	-2	-1

-1	0	1
-2	0	2
-1	0	1

B. Prewitt Operator: The prewitt operator is an approximate way to estimate the magnitude and orientation of an edge. The convolution mask of Prewitt operator is shown in Table 3.

Table 3. Prewitt Mask

1	1	1
0	0	0
-1	-1	-1

-1	0	1
-1	0	1
-1	0	1

C. Roberts Operator: It performs 2-D spatial gradient measurement on an image. It highlights regions of high spatial frequency which often correspond to edges. The cross convolution mask is shown in Table 4

Table 4 Roberts Mask

-1	0	0	1
0	-1	-1	0

D. Laplacian of Guassian (LoG) Operator: It is a second order derivative. The digital implementation of the Laplacian function is made using the mask given in Table 5.

Table 5Laplacian of Guassian (LoG) Operator

0	-1	0
-1	4	-1
0	-1	0

E. Canny Operator: It is a method to find edges by isolating noise from the image without affecting the features of the edges in the image and then applying the tendency to find the edges and the critical value for threshold.

F. Contextual clustering

Image segmentation plays an important role in image analysis and computer vision and it is considered as one of the major obstruction in the development of image processing technology. Recently there has been considerable interest among researchers in statistical clustering techniques in image segmentation was inspired by the methods of statistical physics. These methods were developed to study the equilibrium properties of large, lattice based systems consisting of interacting components as identical. In a clustering technique for image segmentation, each pixel is associated with one of the finite number of categories to form disjoint regions.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The contextual clustering based algorithms are assumed to be drawn from standard normal distribution. It segments a data into category 1 (ω_0) and category 2 (ω_1).

The following are the steps adopted for implementing the contextual clustering algorithm for segmenting the fMRI slice image

(i) Define decision parameter T_{cc} (positive) and weight of neighborhood information β (positive). Let N_n be the total number of data in the neighborhood. Let Z_i be the data itself, i.

(ii) Classify data with $z_i > T_\alpha$ to ω_1 and data to ω_0. Store the classification to C_0 and C_1.

(iii) For each data i, count the number of data u_i, belonging to class ω1 in the neighborhood of data i. Assume that the data outside the range belong to ω_0.

(iv) Classify data with $z_i + \frac{\beta}{T_{cc}}(u_i - \frac{N_n}{2}) > T_\alpha$ to ω_1 and other data to ω_0. Store the classification to variable C_2.

(v) If $C_2 \neq C_1$ and $C_2 \neq C_0$, copy C_1 to C_0, C_2 to C_1 and return to step iii, otherwise stop and return to C_2.

The contextual clustering implementation is as follows:

Step 1: Read a Pattern (fmri image feature).

Step 2: Sort the values of the pattern.

Step 3: Find the Median of the Pattern Cm.

Step 4: Find the number of values greater than the Median Values, U_m.

Step 5: Calculate CC using C_m + (beta/Tcc) * (U_m – (bs/2)).

Step 6: Assign CC as the segmented values.

IV. RESULTS AND DISCUSSION

Earlier researchers had used different metrics to evaluate the segmentation accuracy. In this paper, we have used 'bwlabel' and 'Regionprops' to evaluate the accuracy of segmentation and it has been found that CC segmentation is much better when compared to that of remaining segmentations mentioned in this work.

Figure 1 shows fmri slice. Figures (2-8) show the segmentation by 'Sobel', 'Prewitt', 'Roberts', 'Log', 'Zero crossing', 'Canny', 'CC' methods. Except CC method, in all other segmentation methods, the number of objects are more and , there are some objects segmented are not clear. Matlab 'bwlabel' function has been used and the number objects for each method is shown in Table 6. In addition to .bwlabel', the 'Regionprops' command has been used to find out correct number of segmented objects

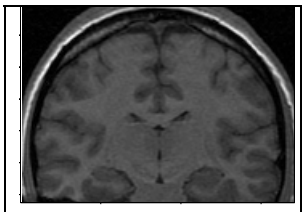


Fig. 1. Sample FMRI image

Fig. 2 Segmentation by 'Sobel' method

Fig 3.Segmentation by' Prewitt' method

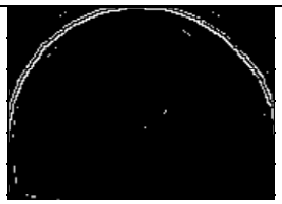


Fig 4.Segmentation by' Roberts' method

Fig. 5 Segmentation by 'Log' method

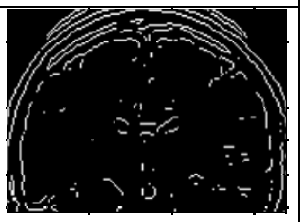


Fig 6.Segmentation by Zero crossing method

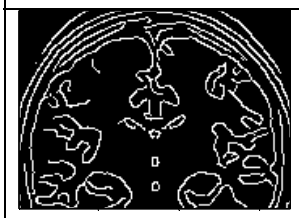


Fig. 7 Segmentation by 'Canny' method

Fig. 8 Segmentation by Contextual clustering

Table 2 Outputs of 'bwlabel' function of 'Matlab 2010'

No	Method	Objects detected
1	Sobel	12
2	Prewitt	11
3	Robertz	8
4	Log	64
5	Zerocross	64

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

6	Canny	32
7	Contextual Clustering	24

V. CONCLUSION

The main purpose of proposing contextual clustering method is to improve segmentation of fMRI images. The supervised contextual clustering extracts features from the fMRI slice that represents information in a given window. The algorithm involves least computation in the segmentation of fMRI slice. The advantages of CC segmentation is that this method uses neighboring information and assured segmentation of minimum one object of fmri image is possible.

REFERENCES

1. Bueno, G., Musse, O., Heitz F, Armspach J.P., 2000, 3D Watershed-based segmentation of internal structures within MR brain images. Medical Images 2000: Image processing, Proc. SPIE, Vol.3979, pp.284-293.
2. Jose, L. Marroquin, Edgar Arce Santana, and Salvador Botello, 2003, Hidden Markov Measure Field Models for Image Segmentation, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol.25, No.11, pp.1380-1387.
3. Liu, S. and Li, J., 2006, Automatic medical image segmentation using gradient and intensity combined level set method, The 28th IEEE Engineering in Medicine and Biological Society, Vol. 1, pp. 3118–3121.
4. Wee, A., Liew, C., and. Yan, H., 2006, Current Methods in the Automatic Tissue Segmentation of 3D Magnetic Resonance Brain Images, Current Medical Imaging Reviews, Vol.2, No.1, pp.91-103.
5. Xiangrong Zhang, Feng Dong, Gordon Clapworthy, Youbing Zhao and Licheng Jiao, 2010, Semi-supervised Tissue Segmentation of 3D Brain MR Images, 14th International Conference Information Visualization, pp.623-628.
6. Yan Li and Zheru Chi, 2005, MR Brain Image Segmentation Based on Self-Organizing Map Network, International Journal of Information Technology, Vol.11, No. 8, pp.45-53.
7. Yongyue Zhang, Michael Brady, and Stephen Smith, 2001, Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm, IEEE Transactions On Medical Imaging, Vol. 20, No. 1, pp.45-57.
8. Zavaljevski, A., Dhawan, A.P., Gaskil, M., Ball, W., Johnson J.D., 2000, Multi-level adaptive segmentation of multi-parameter MR brain images. Comput Med Imag Graphics, Vol. 24, Issue 2, pp. 87–98
9. Zhang, D.Q., Chen, S.C., 2004, A novel kernelized fuzzy c-means algorithm with application in medical image segmentation, Artificial Intelligence Medicine, Vol. 32, Issue 1, pp. 37–50.

	Suganthi D., is pursuing her PhD from Mother Teresa Women's University, Kodikanal, India.
	Dr.S.Purushothaman completed his PhD from Indian Institute of Technology Madras, India in 1995. He has 133 publications to his credit. He has 19 years of teaching experience. Presently he is working as Professor in PET Engineering college , India

IMPLEMENTATION OF HUMAN TRACKING USING BACK PROPAGATION ALGORITHM

[1]Pratheepa S., and [2]Purushothaman S.,

[1]Pratheepa S.,
Research Scholar ,
Mother Teresa Women's University,
Kodaikanal, India-624101.

[2]Dr.S.Purushothaman,
Professor, PET Engineering College,
Vallioor, India-627117.

***ABSTRACT*-Identifying moving objects from a video sequence is a fundamental and critical task in many computer-vision applications. A common approach is to perform background subtraction, which identifies moving objects from the portion of a video frame that differs significantly from a background model. In this work, a new moving object-tracking method is proposed. The moving object is recorded in video. The segmentation of the video is done. Two important properties are used to process the features of the segmented image for highlighting the presence of the human. An artificial neural network with supervised back propagation algorithm learns and provides a better estimate of the movement of the human in the video frame. A multi target human tracking is attempted.**

***Keywords*- Back propagation algorithm (BPA), Human tracking, and Video segmentation**

I. INTRODUCTION

Visual surveillance has become one of the most active research areas in computer vision, especially due to security purposes. Visual surveillance is a general framework that groups a number of different computer vision tasks aiming to detect, track, and classify objects of interests from image sequences, and on the next level to understand and describe these objects behaviors. Haritaoglu, et al, 2000, has stated that tracking people using surveillance equipment has increasingly become a vital tool for many purposes. Among these are the improvement of security and making smarter decisions about logistics and operations of businesses. Automating this process is an ongoing thrust of research in the computer vision community.

Video surveillance systems have long been in use to monitor security sensitive areas. Moving object detection is the basic step for further analysis of video. It handles segmentation of moving objects from stationary background objects. Commonly used techniques for object detection are background subtraction, statistical models, temporal differencing and optical flow. The next step in the video analysis is tracking, which can be defined as the creation of temporal correspondence among detected objects from frame to frame. The output produced by tracking is used to support and enhance motion segmentation, object classification and higher level activity analysis. People tracking is the process of locating a moving people (or many persons) over time using a camera? Tracking people in a video sequence is to determinate the position of the center of gravity and trace the trajectory, or to extract any other relevant information.

II. RELATED WORKS

Beaugendre, et al, 2010 presented efficient and robust object tracking algorithm based on particle filter. The aim is to deal with noisy and bad resolution video surveillance cameras. The main feature used in this method is multi-candidate object detection results based on a background subtraction algorithm combined with color and interaction features. This algorithm only needs a small number of particles to be accurate. Experimental results demonstrate the efficiency of the algorithm for single and multiple object tracking.

Image segmentation's goal is to identify homogeneous region in images as distinct from background and belonging to different objects. A common approach is to classify pixel on the basis of local features (color, position, texture), and then group them together according to their class in order

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

to identify different objects in the scene. For the specific problem of finding moving objects from static cameras, the traditional segmentation approach is to separate pixels into two classes: background and foreground. This is called Background Subtraction [Dengsheng Zhang and Guojun Lu, 2001] and constitutes an active research domain. The output of most background segmentation techniques consists of a bitmap image, where values of 0 and 1 correspond to background and foreground, respectively. Having such a bitmap, the next processing step consists of merging foreground pixels to form bigger groups corresponding to candidate objects; this process is known as object extraction. One common procedure to perform object extraction consists of finding 4 or 8-connected components. This is done using efficient algorithms whose time complexity is linear with respect to the number of pixels in the bitmap. Some of the features used while detecting the moving object such as intensity, color, shape of the region texture, motion in video, display in stereo image, depth in the range Camera temperature in Far infrared, mixture relation between region and stereo disparity. Input video to detect moving object. By using RGB2GRAY predefine function convert colorful video into gray color. The model can be used to detect a moving object in a video. The method generate motion image from consecutive pair of frame. Object is detected in video frames and motion images. From local windows, a neighborhood pixel around background and extract features is formed. Features and background are used to construct and maintain a model, stored in a memory of a computer system.

Hu et al, 2006m proposed a simple and robust method, based on principal axes of people, to match people across multiple cameras. The correspondence likelihood reflecting the similarity of pairs of principal axes of people is constructed according to the relationship between "ground-points" of people detected in each camera view and the intersections of principal axes detected in different camera views and transformed to the same view. The method has the following desirable properties; 1) camera calibration is not needed; 2) accurate motion detection and segmentation are less critical due to the robustness of the principal axis-based feature to noise; 3) based on the fused data derived from correspondence results, positions of people in each camera view can be accurately located even when the people are partially occluded in all views. The experimental results on several real video sequences from outdoor environments have demonstrated the effectiveness, efficiency, and robustness of their method.

Siebel and Stephen, 2002, showed how the output of a number of detection and tracking algorithms can be fused to achieve robust tracking of people in an indoor environment. The new tracking system contains three co-operating parts: i) an Active Shape Tracker using a PCA-generated model of pedestrian outline shapes, ii) a Region Tracker, featuring region splitting and merging for multiple hypothesis matching, and iii) a Head Detector to aid in the initialization of tracks. Data from the three parts are fused together to select the best tracking hypotheses. The new method is validated using sequences from surveillance cameras in an underground station. It is demonstrated that robust real-time tracking of people can be achieved with the new tracking system using standard PC hardware

Martin Spengler and BerntSchiele, 2003, approach is based on the principles of self-organization of the integration mechanism and self-adaptation of the cue models during tracking. Experiments show that the robustness of simple models is leveraged significantly by sensor and model integration.

Tian Hampapur, 2005, proposed a new real-time algorithm to detect salient motion in complex environments by combining temporal difference imaging and a temporal filtered motion field. They assumed that the object with salient motion moves in a consistent direction for a period of time. Compared to background subtraction methods, their method does not need to learn the background model from hundreds of images and can handle quick image variations; e.g., a light being turned on or off. The effectiveness of the proposed algorithm to robust detect salient motion is demonstrated for a variety of real environments with distracting motions.

Zhao and Nevatia, 2004, showed how multiple human objects are segmented and their global motions are tracked in 3D using ellipsoid human shape models. Experiments showed a small number of people move together, have occlusion, and cast shadow or reflection. They estimated the modes e.g., walking, running, standing of the locomotion and 3D body postures by making inference in a prior locomotion model.

Zhuang, et al, 2006, presented a novel approach for visually tracking a colored target in a noisy and dynamic environment using weighted color histogram based particle filter algorithm. In order to make the tracking task robustly and effectively, color histogram based target model is integrated into particle filter algorithm, which considers the target's shape as a necessary factor in target model. Bhattacharyya distance is used to weigh samples in the particle filter by comparing each sample's histogram with a specified target model and it makes the measurement matching and samples' weight updating more reasonable. The method is capable of successfully tracking moving targets in different indoor environment without initial positions information.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. MATERIALS AND METHODOLOGY

A. BACK-PROPAGATION ALGORITHM (BPA)

The BPA uses the steepest-descent method to reach a global minimum. The number of layers and number of nodes in the hidden layers are decided. The connections between nodes are initialized with random weights. A pattern from the training set is presented in the input layer of the network and the error at the output layer is calculated. The error is propagated backwards towards the input layer and the weights are updated. This procedure is repeated for all the training patterns. At the end of each iteration, test patterns are presented to ANN, and the classification performance of ANN is evaluated. Further training of ANN is continued till the desired classification performance is reached.

STEPS INVOLVED.

FORWARD PROPAGATION

The weights and thresholds of the network are initialized.

The inputs and outputs of a pattern are presented to the network.

The output of each node in the successive layers is calculated.

$$o_{(\text{output of a node})} = 1/(1+\exp(-\sum w_{ij}\, x_i + \Theta))$$

The error of a pattern is calculated

$$E(p) = (1/2) \sum (d(p) - o(p))^2$$

REVERSE PROPAGATION

The error for the nodes in the output layer is calculated

$$\delta_{(\text{output layer})} = o(1-o)(d-o)$$

The weights between output layer and hidden layer are updated

$$W_{(n+1)} = W_{(n)} + \eta\delta_{(\text{output layer})}\, o_{(\text{hidden layer})}$$

The error for the nodes in the hidden layer is calculated

$$\delta_{(\text{hidden layer})} = o(1-o) \sum \delta_{(\text{output layer})}\, W_{(\text{updated weights between hidden and output layer})}$$

The weights between hidden and input layer are updated.

$$W_{(n+1)} = W_{(n)} + \eta\delta_{(\text{hidden layer})}\, o_{(\text{input layer})}$$

The above steps complete one weight updation

Second pattern is presented and the above steps are followed for the second weight updation.

When all the training patterns are presented, a cycle of iteration or epoch is completed.

The errors of all the training patterns are calculated and displayed on the monitor as the mean squared error (MSE).

$$E(MSE) = \sum E(p)$$

IV. RESULTS AND DISCUSSION

Identifying all three persons in each frame

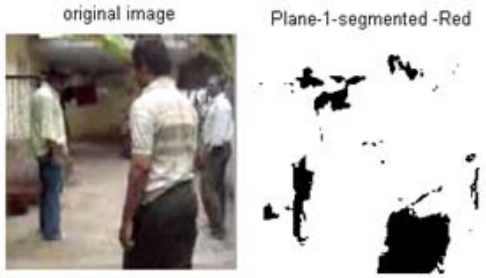


Fig 1 Frame 1 -Original, segmented, video images

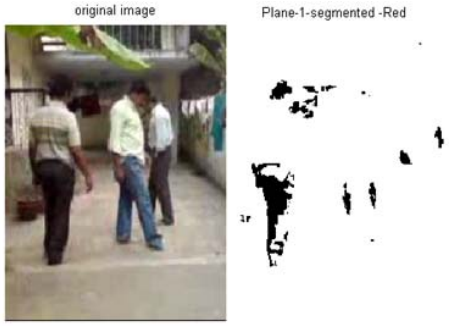


Fig 2.Frame 2 Original, segmented video images

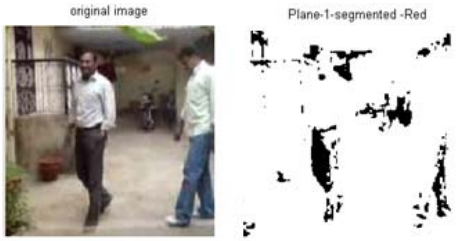


Fig 3 Frame 3 Original, segmented video images

In each frame, three persons are recorded. To segment the frames accurately so that humans are separated from the background irrespective of varied illumination. Figures 1 to 3 identify the different

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

persons in a frame and track the same persons in subsequent frames. Figure 4 presents the estimation by BPA.

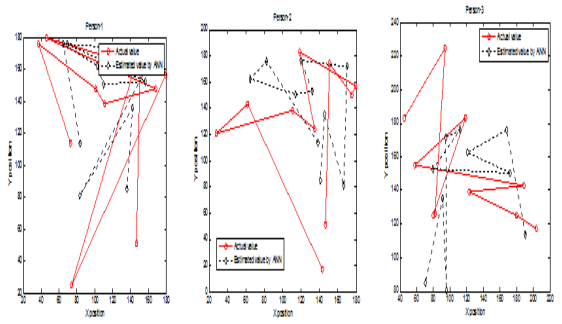


Fig.4 Implementation of back propagation neural network for tracking three persons in video frame

An artificial neural network with supervised back propagation algorithm learns the input output scenarios and provides a better estimate of the movement of the human in the video frame. A multi target human tracking is attempted.

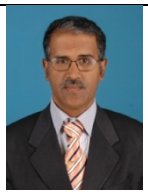
V. CONCLUSION

Video tracking is an important process in tracking humans. It involves various image processing concepts. In this work, the acquired video has been separated into frames and segmented. From the segmented frames, the humans are identified and compared with template. Based on the comparisons, the human is tracked.

REFERENCES

1. Beaugendre A., Miyano, H., shidera E., Goto S., 2010, Human tracking system for automatic video surveillance with particle filters, IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), pp.152-155.
2. Dengsheng Zhang and Guojun Lu, 2001, Segmentation in moving object in image sequence: A review, Circuit system signal processing, Vol.20, N0.2, pp.143-183.
3. Haritaoglu I., Harwood D., Davis L.S., 2000, W4: Real-Time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.8, pp.809-830.
4. Hu W., Min Hu., Xue Zhou ,Tan T.N., 2006, Principal axis-based correspondence between multiple cameras for people tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.28, No.4, pp.663-671.
5. Nils T. Siebel, and Stephen J. Maybank, 2002, Fusion of Multiple Tracking Algorithms for Robust People Tracking, European Conference on Computer Vision, pp.373–387.
6. Spengler M., and Schiele B., 2003, Towards robust multi-cue integration for visual tracking, Machine Vision and Applications, Vol.14, pp.50–58.
7. Tian Y.L., and Hampapur A., 2005, Robust Salient Motion Detection with Complex Background for Real-time Video Surveillance, IEEE Computer Society Workshop on Motion and Video Computing, Vol.2, pp.30–35.
8. Zhao T., Nevatia R., 2004, Tracking Multiple Humans in Complex Situations, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.26, Issue 9, pp.1208–1221.
9. Zhuang Y., Wang W., and Xing R.Z., 2006, Target Tracking in Colored Image Sequence Using Weighted Color Histogram Based Particle Filter, IEEE International Conference on Robotics and Biometrics, pp.1488-1493.

S.Pratheepa completed her M.Phil from Manonamaniam Sundaranar university, India in 2003. Now She is doing Ph.D. in Mother Teresa Women's University, Kodaikanal. She has 11 years of teaching experience. Presently she is working as Head & Asst. Professor in J.H.A Agarsen College, Chennai.



Dr.S.Purushothaman completed his PhD from Indian Institute of Technology Madras, India in 1995. He has 133 publications to his credit. He has 19 years of teaching experience. Presently he is working as Professor in PET Engineering college , India

Stock Market Trend Analysis Using Hidden Markov Models

Kavitha G
School of Applied Sciences,
Hindustan University, Chennai, India.

***Udhayakumar A**
School of Computing Sciences,
Hindustan University, Chennai, India

Nagarajan D
Department of Information Technology, Math Section,
Salalah College of Technology,
Salalah, Sultanate of Oman

Abstract — Price movements of stock market are not totally random. In fact, what drives the financial market and what pattern financial time series follows have long been the interest that attracts economists, mathematicians and most recently computer scientists [17]. This paper gives an idea about the trend analysis of stock market behaviour using Hidden Markov Model (HMM). The trend once followed over a particular period will sure repeat in future. The one day difference in close value of stocks for a certain period is found and its corresponding steady state probability distribution values are determined. The pattern of the stock market behaviour is then decided based on these probability values for a particular time. The goal is to figure out the hidden state sequence given the observation sequence so that the trend can be analyzed using the steady state probability distribution(π) values. Six optimal hidden state sequences are generated and compared. The one day difference in close value when considered is found to give the best optimum state sequence.

Keywords-Hidden Markov Model; Stock market trend; Transition Probability Matrix; Emission Probability Matrix; Steady State Probability distribution

I. INTRODUCTION

"A growing economy consists of prices falling, not rising", says Kel Kelly[9]**.** Stock prices change every day as a result of market forces. There is a change in share price because of supply and demand. According to the supply and demand, the stock price either moves up or undergoes a fall. Stock markets normally reflect the business cycle of the economy: when the economy grows, the stock market typically reflects this economic growth in an upward trend in prices. In contrast, when the economy slows, stock prices tend to be more mixed. Markets may take time to form bottoms or make tops, sometimes of two years or more. This makes it difficult to determine when the market hits a top or a bottom[3]. The Stock Market patterns are non-linear in nature, hence it is difficult to forecast future trends of the market behaviour.

*Corresponding author

In this paper, a method has been developed to forecast the future trends of the stock market. The Latent or hidden states, which determine the behaviour of the stock value, are usually invisible to the investor. These hidden states are derived from the emitted symbols. The emission probability depends on the current state of the HMM. Probability and Hidden Markov Model give a way of dealing with uncertainty. Many intelligent tasks are sequence finding tasks, with a limited availability of information. This naturally involves hidden states or strategies for dealing with uncertainty.

II. LITERATURE SURVEY

In Recent years, a variety of forecasting methods have been proposed and implemented for the stock market analysis. A brief study on the literature survey is presented. Markov Process is a stochastic process where the probability at one time is only conditioned on a finite history, being in a certain state at a certain time. Markov chain is "Given the present, the future is independent of the past". HMM is a form of probabilistic finite state system where the actual states are not directly observable. They can only be estimated using observable symbols associated with the hidden states. At each time point, the HMM emits a symbol and changes a state with certain probability. HMM analyze and predict time series or time depending phenomena. There is not a one to one correspondence between the states and the observation symbols. Many states are mapped to one symbol and vice-versa.

Hidden Markov Model was first invented in speech recognition [12,13], but is widely applied to forecast stock market data. Other statistical tools are also available to make forecasts on past time series data. Box–Jenkins[2] used Time series analysis for forecasting and control. White[5,18,19] used Neural Networks for stock market forecasting of IBM daily stock returns. Following this, various studies reported on the effectiveness of alternative learning algorithms and prediction methods using ANN. To forecast the daily close and morning open price, Henry [6] used ARIMA model. But all these conventional methods had problems when non linearity exists in time series. Chiang et al.[4] have used ANN to forecast the end-of-year net asset value of mutual funds. Kim and Han [10] found that the complex dimensionality and buried

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

noise of the stock market data makes it difficult to re-estimate the ANN parameters. Romahi and Shen [14] also found that ANN occasionally suffers from over fitting problem. They developed an evolving rule based expert system and obtained a method which is used to forecast financial market behaviour. There were also hybridization models effectively used to forecast financial behaviour. The drawback was requirement of expert knowledge. To overcome all these problems Hassan and Nath [15] used HMM for a better optimization. Hassan et al. [16] proposed a fusion model of HMM, ANN and GA for stock Market forecasting. In continuation of this, Hassan [7] combined HMM and fuzzy logic rules to improve the prediction accuracy on non-stationary stock data sets. Jyoti Badge[8] used technical indicators as an input variable instead of stock prices for analysis. Aditya Gupta and Bhuwan Dhingra[1] considered the fractional change in Stock value and the intra-day high and low values of the stock to train the continuous HMM**.** In the earlier studies, much research work had been carried out using various techniques and algorithms for training the model for forecasting or predicting the next day close value of the stock market, for which randomly generated Transition Probability Matrix (TPM), Emission Probability Matrix (EPM) and prior probability matrix have been considered.

In this paper, the trend analysis of the stock market is found using Hidden Markov Model by considering the one day difference in close value for a particular period. For a given observation sequence, the hidden sequence of states and their corresponding probability values are found. The probability values of π gives the trend percentage of the stock prices. Decision makers make decisions in case of uncertainty. The proposed approach gives a platform for decision makers to make decisions on the basis of the percentage of probability values obtained from the steady state probability distribution.

III. RESEARCH SET UP

A. Basics of HMM

HMM is a stochastic model where the system is assumed to be a Markov Process with hidden states. HMM gives better accuracy than other models. Using the given input values, the parameters of the HMM (λ) denoted by A, B and π are found out.

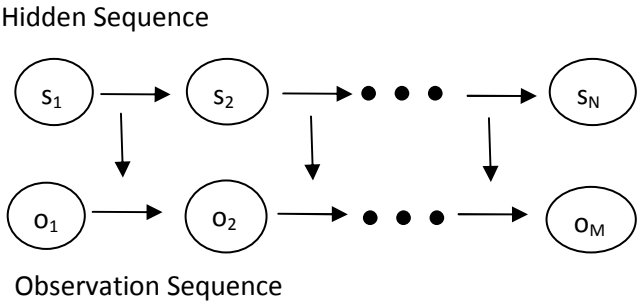


Fig 1. **Trellis Diagram**

HMM consists of

A set of hidden or latent states (S)

A set of possible output symbols (O)

A state transition probability matrix (A)

- probability of making transition from one state to each of the other states

Observation emission probability matrix (B)

- probability of emitting/observing a symbol at a particular state

Prior probability matrix (π)

- probability of starting at a particular state

An HMM is defined as $\lambda=(S, O, A, B, \pi)$ where

$S=\{s_1,s_2,\ldots,s_N\}$ is a set of N possible states

$O=\{o_1,o_2,\ldots,o_M\}$ is a set of M possible observation symbols

A is an NxN state Transition Probability Matrix (TPM)

B is an NxM observation or Emission Probability Matrix (EPM)

π is an N dimensional initial state probability distribution vector

and A, B and π should satisfy the following conditions:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \text{where} \quad 1 \le i \le N;$$

$$\sum_{j=1}^{M} b_{ij} = 1 \quad \text{where} \quad 1 \le i \le N;$$

$$\sum_{i=1}^{N} \pi_i = 1 \quad \text{where} \quad \pi_i \ge 0$$

The main problems of HMM are: Evaluation, Decoding, and Learning.

Evaluation problem

Given the HMM $\lambda = \{A, B, \pi\}$ and the observation sequence $O=o_1\ o_2\ ...\ o_M$, the probability that model λ has generated sequence O is calculated.

Often this problem is solved by the Forward Backward Algorithm (Rabiner, 1989) (Rabiner, 1993).

Decoding problem

Given the HMM $\lambda = \{A, B, \pi\}$ and the observation sequence $O=o_1\ o_2\ ...\ o_M$, calculate the most likely sequence of hidden states that produced this observation sequence O.

Usually this problem is handled by Viterbi Algorithm (Rabiner,1989) (Rabiner,1993).

Learning problem

Given some training observation sequences $O=o_1\ o_2\ ...\ o_M$ and general structure of HMM (numbers of hidden and visible states), determine HMM parameters $\lambda = \{A, B, \pi\}$ that best fit training data.

The most common solution for this problem is Baum-Welch algorithm (Rabiner,1989) (Rabiner,1993) which is considered as the traditional method for training HMM.

In this paper, IBM daily close value data for a month period is considered.

Two observing symbols " I " and " D " have been used:

"I indicates Increasing" , " D indicates Decreasing ".

If Today's close value – Yesterday's close value > 0, then observing symbol is I

If Today's close value – Yesterday's close value < 0 then observing symbol is D

There are six hidden states assumed and are denoted by the symbol

S1, S2 , S3 , S4, S5, S6

where

S1 indicates "very low";
S2 indicates "low";
S3 indicates "moderate low"
S4 indicates "moderate high";
S5 indicates "high";
S6 indicates "very high".

The States are not directly observable. The situations of the stock market are considered hidden. Given a sequence of observation we can find the hidden state sequence that produced those observations.

B. Database

The complete set of data for the proposed study has been taken from *yahoofinance.com.* The Table 1 given below shows the daily close value of the stock market:

Table I. Daily close value for finding differences in one day, two days, three days, four days, five days, six days close value

S.NO	C.V	D.in.1 day CV	O.S	D.in.2 days CV	O.S	D.in.3 days CV	O.S	D.in4 days CV	O.S	D.in.5 days CV	O.S	D.in.6 days CV	O.S
1	77.91												
2	77.39	-0.52	D										
3	76.5	-0.89	D	-1.41	D								
4	75.86	-0.64	D	-1.53	D	-2.05	D						
5	77.45	1.59	I	0.95	I	0.06	I	-0.46	D				
6	79.33	1.88	I	3.47	I	2.83	I	1.94	I	1.42	I		
7	79.51	0.18	I	2.06	I	3.65	I	3.01	I	2.12	I	1.6	I
8	79.15	-0.36	D	-0.18	D	1.7	I	3.29	I	2.65	I	1.76	I
9	79.95	0.8	I	0.44	I	0.62	I	2.5	I	4.09	I	3.45	I
10	78.56	-1.39	D	-0.59	D	-0.95	D	-0.77	D	1.11	I	2.7	I
11	79.07	0.51	I	-0.88	D	-0.08	D	-0.44	D	-0.26	D	1.62	I
12	77.4	-1.67	D	-1.16	D	-2.55	D	-1.75	D	-2.11	D	-1.93	D
13	77.28	-0.12	D	-1.79	D	-1.28	D	-2.67	D	-1.87	D	-2.23	D
14	77.95	0.67	I	0.55	I	-1.12	D	-0.61	D	-2	D	-1.2	D
15	77.33	-0.62	D	0.05	I	-0.07	D	-1.74	D	-1.23	D	-2.62	D
16	76.7	-0.63	D	-1.25	D	-0.58	D	-0.7	D	-2.37	D	-1.86	D
17	77.73	1.03	I	0.4	I	-0.22	D	0.45	I	0.33	I	-1.34	D
18	77.07	-0.66	D	0.37	I	-0.26	D	-0.88	D	-0.21	D	-0.33	D
19	77.9	0.83	I	0.17	I	1.2	I	0.57	I	-0.05	D	0.62	I
20	75.7	-2.2	D	-1.37	D	-2.03	D	-1	D	-1.63	D	-2.25	D

C.V – Close value ; O.S – Observing symbol

D.in.1 day CV - difference in 1 day close value;
D.in.2 days CV - difference in 2 days close value;
D.in.3 days CV - difference in 3 days close value;
D.in.4 days CV - difference in 4 days close value;
D.in.5 days CV - difference in 5 days close value;
D.in.6 days CV - difference in 6 days close value

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. CALCULATION

The various probability values of TPM, EPM and π for difference in one day, two days, three days, four days, five days, six days close value close value are calculated as given below.

A. *Probability values of TPM, EPM and π for difference in one day close value:*

	S1	S2	S3	S4	S5	S6
S1	0	0	1	0	0	0
S2	0	0	0.5	0.5	0	0
S3	0	0.143	0.143	0	0.571	0.143
S4	0.5	0	0.5	0	0	0
S5	0.25	0.25	0.5	0	0	0
S6	0	0	0	0.5	0	0.5

Fig 2. **TPM**

	I	D
S1	0	1
S2	0.5	0.5
S3	0.71	0.29
S4	0	1
S5	0	1
S6	1	0

Fig 3. **EPM**

Steady state probability distribution

$\pi = [0.06 \quad 0.11 \quad 0.39 \quad 0.11 \quad 0.22 \quad 0.11]$

Table II. Transition table with probability values for difference in one day close value

TRANSITION OF STATES WITH OBSERVING SYMBOLS	S1		S2		S3		S4		S5		S6	
	I	D	I	D	I	D	I	D	I	D	I	D
S1	0	0	0	0	0	1	0	0	0	0	0	0
S2	0	0	0	0	0	0.5	0.5	0	0	0	0	0
S3	0	0	0	0.1429	0	0.1429	0	0	0.5714	0	0.1429	0
S4	0	0.5	0	0	0	0.5	0	0	0	0	0	0
S5	0	0.25	0	0.25	0	0.5	0	0	0	0	0	0
S6	0	0	0	0	0	0	0	0	0	0	1	0

B. *Probability values of TPM, EPM and π for difference in two days close value:*

	S1	S2	S3	S4	S5	S6
S1	0.4	0	0.4	0.2	0	0
S2	0.33	0.33	0.33	0	0	0
S3	0.33	0.17	0.5	0	0	0
S4	0	0	0	0	0	1
S5	0	1	0	0	0	0
S6	0	0	0	0	1	0

Fig 4. **TPM**

	I	D
S1	0.6	0.4
S2	0.33	0.67
S3	0.5	0.5
S4	1	0
S5	0	1
S6	0	1

Fig 5. **EPM**

Steady state probability distribution

$\pi = [0.29 \quad 0.18 \quad 0.35 \quad 0.06 \quad 0.06 \quad 0.06]$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table III. Transition table with probability values for difference in two days close values

TRANSITION OF STATES WITH OBSERVING SYMBOLS	S1		S2		S3		S4		S5		S6	
	I	D	I	D	I	D	I	D	I	D	I	D
S1	0	0.4	0	0	0.4	0	0.2	0	0	0	0	0
S2	0	0.33	0	0.33	0.33	0	0	0	0	0	0	0
S3	0	0.33	0	0.167	0.5	0	0	0	0	0	0	0
S4	0	0	0	0	0	0	0	0	0	0	1	0
S5	0	0	0	1	0	0	0	0	0	0	0	0
S6	0	0	0	0	0	0	0	0	0	1	0	0

C. Probability values of TPM, EPM and π for difference in three days close value:

$$\begin{array}{c} \\ S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{array} \begin{array}{c} \begin{array}{cccccc} S1 & S2 & S3 & S4 & S5 & S6 \end{array} \\ \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0.25 & 0.75 & 0 & 0 & 0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0.2 \\ 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix} \end{array}$$

Fig 6. **TPM**

$$\begin{array}{c} \\ S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{array} \begin{array}{c} \begin{array}{cc} I & D \end{array} \\ \begin{bmatrix} 0.5 & 0.5 \\ 0 & 1 \\ 0.4 & 0.6 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \end{array}$$

Fig 7. **EPM**

Steady state probability distribution

$\pi = [0.13 \quad 0.25 \quad 0.31 \quad 0.13 \quad 0.06 \quad 0.13]$

Table 1V. Transition table with probability values for difference in three days close values

TRANSITION OF STATES WITH OBSERVING SYMBOLS	S1		S2		S3		S4		S5		S6	
	I	D	I	D	I	D	I	D	I	D	I	D
S1	0	0	0	0.5	0.5	0	0	0	0	0	0	0
S2	0	0	0	0.25	0	0.75	0	0	0	0	0	0
S3	0	0.2	0	0.2	0	0.2	0.2	0	0	0	0.2	0
S4	0	0.5	0	0.5	0	0	0	0	0	0	0	0
S5	0	0	0	0	0	0	1	0	0	0	0	0
S6	0	0	0	0	0	0	0	0	0.5	0	0.5	0

D. Probability values of TPM, EPM and π for difference in four days close value:

$$\begin{array}{c} \\ S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{array} \begin{array}{c} \begin{array}{cccccc} S1 & S2 & S3 & S4 & S5 & S6 \end{array} \\ \begin{bmatrix} 0.33 & 0.33 & 0.33 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0.67 & 0 & 0 \\ 0.67 & 0 & 0 & 0 & 0.33 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0.33 & 0 & 0 & 0 & 0.67 \end{bmatrix} \end{array}$$

Fig 8. **TPM**

$$\begin{array}{c} \\ S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{array} \begin{array}{c} \begin{array}{cc} I & D \end{array} \\ \begin{bmatrix} 0 & 1 \\ 0.67 & 0.33 \\ 0.33 & 0.67 \\ 0 & 1 \\ 1 & 0 \\ 0.67 & 0.33 \end{bmatrix} \end{array}$$

Fig 9. **EPM**

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Steady state probability distribution

$\pi = [0.04 \quad 0.04 \quad 0.04 \quad 0.13 \quad 0.07 \quad 0.04]$

Table V. Transition table with probability values for difference in four days close values

TRANSITION OF STATES WITH OBSERVING SYMBOLS	S1		S2		S3		S4		S5		S6	
	I	D	I	D	I	D	I	D	I	D	I	D
S1	0	0.33	0	0.33	0	0.33	0	0	0	0	0	0
S2	0	0	0	0	0	0.33	0.67	0	0	0	0	0
S3	0	0.67	0	0	0	0	0	0	0.33	0	0	0
S4	0	0	0	1	0	0	0	0	0	0	0	0
S5	0	0	0	0	0	0	0	0	0	0	1	0
S6	0	0	0	0.33	0	0	0	0	0	0	0.67	0

E. *Probability values of TPM, EPM and π for difference in five days close value:*

	S1	S2	S3	S4	S5	S6
S1	0.5	0.25	0.25	0	0	0
S2	1	0	0	0	0	0
S3	0.33	0	0.67	0	0	0
S4	0	0.5	0	0	0.5	0
S5	0	0	0	0	0.5	0.5
S6	0	0	0	1	0	0

Fig 10. **TPM**

	I	D
S1	0.25	0.75
S2	0	1
S3	0	1
S4	0.5	0.5
S5	1	0
S6	1	0

Fig 11. **EPM**

Steady state probability distribution

$\pi = [0.29 \quad 0.14 \quad 0.21 \quad 0.14 \quad 0.14 \quad 0.07]$

Table VI. Transition table with probability values for difference in five days close values

TRANSITION OF STATES WITH OBSERVING SYMBOLS	S1		S2		S3		S4		S5		S6	
	I	D	I	D	I	D	I	D	I	D	I	D
S1	0.5	0	0	0.25	0.25	0	0	0	0	0	0	0
S2	0	1	0	0	0	0	0	0	0	0	0	0
S3	0	0.33	0	0	0	0.67	0	0	0	0	0	0
S4	0	0	0	0.5	0	0	0	0	0.5	0	0	0
S5	0	0	0	0	0	0	0	0	0.5	0	0.5	0
S6	0	0	0	0	0	0	1	0	0	0	0	0

F. *Probability values of TPM, EPM and π for difference in six days close value:*

	S1	S2	S3	S4	S5	S6
S1	0.5	0.5	0	0	0	0
S2	0.5	0	0.5	0	0	0
S3	0	0	0	1	0	0
S4	1	0	0	0	0	0
S5	0.33	0	0	0	0.33	0.33
S6	0	0	0	0	0.5	0.5

Fig 12. **TPM**

	I	D
S1	0	1
S2	0	1
S3	1	0
S4	0	1
S5	0.67	0.33
S6	1	0

Fig 13. **EPM**

Steady state probability distribution

$\pi = [0.31 \quad 0.15 \quad 0.08 \quad 0.08 \quad 0.23 \quad 0.15]$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table VII. **Transition table with probability values for difference in six days close values**

TRANSITION OF STATES WITH OBSERVING SYMBOLS	S1		S2		S3		S4		S5		S6	
	I	D	I	D	I	D	I	D	I	D	I	D
S1	0	0.5	0	0.5	0	0	0	0	0	0	0	0
S2	0	0.5	0	0	0	0.5	0	0	0	0	0	0
S3	0	0	0	0	0	0	1	0	0	0	0	0
S4	0	1	0	0	0	0	0	0	0	0	0	0
S5	0	0.33	0	0	0	0	0	0	0.33	0	0.33	0
S6	0	0	0	0	0	0	0	0	0.5	0	0.5	0

The MATLAB function "Hmmgenerate" is used to generate a random sequence of emission symbols and states. The length of both sequence and states to be generated is denoted by L.

The HMM matlab toolbox syntax is :

[Sequence, States] = Hmmgenerate (L , TPM, EPM) , see [11]

For instance,

If the Input is given as,

TPM = [0 0 1 0 0 0; 0 0 0.5 0.5 0 0; 0 0.143 0.143 0 0.571 0.143; 0.5 0 0.5 0 0 0; 0.25 0.25 0.5 0 0 0; 0 0 0 0.5 0 0.5];

EPM = [0 1;0.5 0.5; 0.71 0.29;0 1;0 1;1 0];

[sequence,states] = hmmgenerate(7, TPM, EPM)

```
'Sequence Symbols',{'I','D'},...
    'Statenames',{'very
low';'low';'moderate low';'moderate
high';'high';'very high'}
```

Then the Output of few randomly generated sequences and states is given below:

Sequence: $\mathcal{E}$ → I → D → D → I → I → I → I

states : S3 S2 S3 S6 S6 S6 S6

sequence : $\mathcal{E}$ → D → I → D → D → I → I → I

states : S3 S3 S5 S1 S3 S2 S3

where '$\mathcal{E}$' denotes the start symbol .

The fitness function used for finding the fitness value of sequence of states is defined by

$$\text{Fitness} = \frac{1}{\sum compare(i, j)}$$

V. DISCUSSION

Using the Iterative procedure, for each TPM and EPM framed we get an optimum sequence of states generated. The length of the sequence generated is taken as L=7, for instance.

The optimum sequence of states obtained from the one day difference TPM and EPM is

1. $\mathcal{E}$ → D → I → D → I → D → I
   S1 S3 S5 S3 S5 S3 S5

Similarly ,we get 5 more such optimum sequences of states for 2 day difference , 3 day difference, 4 day difference, 5 day difference, 6 day difference TPM and EPM respectively as follows:

2. $\mathcal{E}$ → I → D → D → I → D → D
   S1 S3 S1 S1 S3 S1 S1

3. $\mathcal{E}$ → D → D → I → D → I → I
   S1 S2 S3 S4 S1 S3 S4

4. $\mathcal{E}$ → D → I → D → I → D → D
   S1 S2 S4 S2 S4 S2 S3

5. $\mathcal{E}$ → D → D → I → I → D → D
   S1 S2 S1 S1 S1 S2 S1

6. $\mathcal{E}$ → D → D → I → D → D → D
   S1 S2 S3 S4 S1 S2 S3

Using the fitness function we compute the fitness value for each of the optimum sequence of states obtained.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table VIII. Comparison of Six Optimum State Sequences

S.No.	Comparison of 6 optimum sequence of states	Calculated value	Fitness = $\frac{1}{\sum compare(i,j)}$
1.	(1,2) + (1,3) + (1,4) + (1,5) + (1,6)	1	1
2.	(2,1) + (2,3) + (2,4) + (2,5) + (2,6)	1.29	0.76
3.	(3,1) + (3,2) + (3,4) + (3,5) + (3,6)	1.86	0.54
4.	(4,1) + (4,2) + (4,3) + (4,5) + (4,6)	1.43	0.70
5.	(5,1) + (5,2) + (5,3) + (5,4) + (5,6)	2.14	0.47
6.	(6,1) + (6,2) + (6,3) + (6,4) + (6,5)	2.14	0.47

VI. CONCLUSION

In this paper, results are presented using Hidden Markov Model to find the trend of the stock market behavior. The highest is the fitness value, the better is the performance of the particular sequence. One day difference in close value when considered is found to give the best optimum sequence. It is observed that at any point of time over years, if the stock market behaviour pattern is the same then we can observe the same steady state probability values as obtained in one day difference of close value, which clearly determines the behavioural pattern of the stock market.

VII. REFERENCES

[1] Aditya Gupta and Bhuwan Dhingra, Non-Student members, IEEE, "Stock Market Prediction Using Hidden Markov Models," 2012.

[2] G. E. P. Box and G. M. Jenkins, Time series analysis: forecasting and control. San Fransisco, CA: Holden-Day, 1976.

[3] Dr. Bryan Taylor, The Global Financial Data Guide to Bull and Bear Markets, President, Global Financial Data, Inc.

[4] W.C. Chiang, T. L. Urban and G. W. Baldridge,"A neural network approach to mutual fund net asset value forecasting," Omega International Journal of Management Science.,Vol. **24** (2), pp. 205–215,1996.

[5] Halbert White, "Economic prediction using neural networks: the case of IBM daily stock returns," Department of Economics, University of California, San Diego.

[6] Henry M. K. Mok, "Causality of interest rate, exchange rate and stock prices at stock market open and close in Hong Kong," Asia Pacific Journal of Management,Vol. **10** (2), pp. 123–143,1993.

[7] Hassan Rafiul, Nath Baikunth and Michael Kirley, "HMM based Fuzzy Model for Time Series Prediction," IEEE International Conference on Fuzzy Systems., pp. 2120-2126,2006.

[8] Jyoti Badge, "Forecasting of Indian Stock Market by Effective Macro- Economic Factors and Stochastic Model," Journal of Statistical and Econometric Methods, Vol. **1** (2),pp. 39-51,ISSN: 2241-0384 (print), 2241-0376 (online) Sciencepress Ltd, 2012.

[9] Kel Kelly, A growing economy consists of prices falling, not rising, How the Stock Market and Economy Really Work , 2010.

[10] K.J. Kim and I. Han, "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, Expert Systems with Applications," Vol.19,pp. 125–132,2000.

[11] K. Murphy, HMM Toolbox for MATLAB, Internet: http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm .html, Oct. 29, 2011.

[12] L. Rabiner and B. Juang, "Fundamentals of Speech Recognition," Prentice-Hall, Englewood Cliffs,NJ, 1993.

[13] L.R Rabiner, "A tutorial on HMM and Selected Applications in Speech Recognition," In:[WL], proceedings of the IEEE,Vol. **77** (2), pp. 267-296,1993.

[14] Y. Romahi and Q. Shen, "Dynamic financial forecasting with automatically induced fuzzy associations," In Proceedings of the 9th international conference on fuzzy systems., pp. 493–498,2000.

[15] Md. Rafiul Hassan and Baikunth Nath, "Stock Market forecasting using Hidden Markov Model: A New Approach," Proceeding of the 2005 5th international conference on intelligent Systems Design and Application 0-7695-2286-06/05, IEEE, 2005.

[16] Md. Rafiul Hassan, Baikunth Nath and Michael Kirley, " A fusion model of HMM, ANN and GA for stock market forecasting," Expert systems with Applications., pp. 171-180,2007.

[17] A. S. Weigend A. D. Back, "What Drives Stock Returns?-An Independent Component Analysis," In Proceedings of the IEEE/IAFE/INFORMS 1998 Conference on Computational Intelligence for Financial Engineering, IEEE, New York., pp. 141-156,1998.

[18] H.White, "Economic prediction using neural networks: the case of IBM daily stock returns," In Proceedings of the second IEEE annual conference on neural networks., II, pp. 451–458,1988.

[19] H.White, Learning in artificial neural networks: a statistical perspective, Neural Computation., Vol. **1** , pp. 425–464,1989.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Computational impact of hydrophobicity in protein stability

Geetika S. Pandey[1]
Research Scholar,
CSE dept., RGPV,
Bhopal (M.P), India

Dr. R.C Jain[2]
Director, SATI(D),
Vidisha(M.P), India

***Abstract*- Among the various features of amino acids, the hydrophobic property has most visible impact on stability of a sequence folding. This is mentioned in many protein folding related work, in this paper we more elaborately discuss the computational impact of the well defined 'hydrophobic aspect in determining stability', approach with the help of a developed 'free energy computing algorithm' covering various aspects - preprocessing of an amino acid sequence, generating the folding and calculating free energy. Later discussing its use in protein structure related research work.**

***Keywords*- *amino acids, hydrophobicity, free energy, protein stability*.**

I. INTRODUCTION

Since the earliest of proteomics researches, it has been clear that the positioning and properties of amino acids are key to structural analysis [1]. According to Betts et.al. in the protein environment a feature of key importance is cellular location. Different parts of cells have very different chemical environments with the consequence that many amino acids behave differently. The biggest difference as mentioned by Betts et.al. is between soluble proteins and membrane proteins. The soluble proteins tend to be surrounded by water molecules i.e have polar or hydrophilic residues on their surface whereas membrane proteins are surrounded by lipids i.e they tend to have hydrophobic residues on the surface that interact with the membrane. Further the soluble proteins are categorized as extracellular and intracellular. So basically through the various studies [2] could conclude that the core of protein contains hydrophobic amino acids forming certain bonds and thus structures. the stability of the structures is determined by the free energy change , as mentioned by Zhang et. al [3] i.e.

$\Delta G(folding)=G(folded)-G(unfolded)$ [3]

Later in this paper various aspects of folding and stability are discussed in detail.

II. BACKGROUND

A. Features

Shaolei Teng et.al.[4] mentioned twenty amino acid features which they used to code each amino acid residue in a data instance. They obtained these features from Protscale (http://expasy.org/tools/protscale.html) [5] and AAindex (http://www.genome.jp/aaindex/) [6]. They further mentioned these features into four categories -

Biochemical features – includes M, molecular weight, this is related to volume of space that a residue occupies in protein structure. K, side chain pka value, which is related to the ionization state of a residue and thus plays a key role in pH dependent protein stability. H, hydrophobicity index, which is important for amino acid side chain packing and protein folding. The hydrophobic interactions make non-polar side chains to pack together inside proteins

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

and disruption of these interactions may cause protein destabilization. P, polarity, which is the dipole-dipole intermolecular interactions between the positively and negatively charged residues. Co, overall amino acid composition, which is related to the evolution and stability of small proteins.

Structural features- this includes A, alpha-helix. B, beta-sheet. C, coil. Aa, average area buried on transfer from standard state to folded protein. Bu, bulkiness, the ratio of the side chain volume to the length of the amino acid.

Empirical Features- this includes, S1, protein stability scale based on atom atom potential of mean force based on Distance Scaled Finite Ideal-gas Reference (DFIRE). S2, relative protein stability scale derived from mutation experiments. S3, side-chain contribution to protein stability based on data from protein denaturation experiments.

Other biological features- F, average flexibility index. Mc, mobility of an amino acid on chromatography paper. No, number of codons for an amino acid. R, refractivity, protein density and folding characteristics. Rf, recognition factor, average of stabilization energy for an amino acid. Rm, relative mutability of an amino acid. Relative mutability indicates the probability that a given amino acid can be changed to others during evolution. Tt, transmembrane tendency scale. F, average flexibility index of an amino acid derived from structures of globular proteins.

B. Protein folding

Protein folding has been considered as one of the most important process in biology. under the various physical and chemical conditions the protein sequences fold forming bonds , when these conditions are favourable the folding leads to proper biological functionality. But some conditions could lead to denaturation of the structures thus giving unfolded structures. protein denaturants could be [7] –

- High temperatures, can cause protein unfolding, aggregation.
- Low temperatures, some proteins are sensitive to cold denaturation.
- Heavy metals(e.g. lead, cadmium etc), highly toxic, efficiently induce the 'stress response'.
- Proteotoxic agents(e.g. alcoholc, cross-linking agents etc.)
- Oxygen radicals, ionizing radiation- can cause permanent protein damage.
- Chaotropes (urea, guandine hydrochloride etc.), highly potent at denaturing proteins, often used in protein folding studies.

Protein folding considers the question of how the process of protein folding occurs, i.e how the unfolded protein adopts the native state. Very often this problem has been described as the second half of the genetic code. Studies till date conclude the following steps as the solution for this problem [8] –

- 3D structure prediction from primary sequence.
- Avoiding misfolding related to human diseases.
- Designing proteins with novel functions.

C. Factors affecting protein stability

Protein stability is the net balance of forces which determine whether a protein will be in its native folded conformation or a denatured state. Negative enthalpy change and positive entropy change give negative i.e. stabilizing, contributions to the free energy of protein folding, i.e. the lower the ΔG, the more stable the protein structure is [7]. Any situation that minimizes the area of contact between H_2O and non-polar, i.e hydrocarbon, regions of the protein results in an increase in entropy [9].

$$\Delta G = \Delta H - T\Delta S$$

Following are the factors affecting protein stability [8]:

- pH : proteins are most stable in the vicinity of their isoelectric point, pI. In general, with some exceptions, electrostatic interactions are believed to contribute to a small amount of the stability of the native state.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Ligand binding: binding ligands like inhibitors to enzymes, increases the stability of the protein.
- Disulphide bonds: it has been observed that many extracellular proteins contained disulphide bonds, whereas intracellular proteins usually did not exhibit disulphide bonds. Disulphide bonds are believed to increase the stability of the native state by decreasing the conformational entropy of the unfolded state due to the conformational constraints imposed by cross linking (i.e decreasing the entropy of the unfolded state).
- Dissimilar properties of residues: not all residues make equal contributions to protein stability. Infact, studies say that the interior ones, inaccessible to the solvent in the native state make a much greater contribution than those on the surface.

III. EXPERIMENTAL PROCEDURE

A. *Approach*

As per the amino acid features mentioned previously, the hydrophobic property is most responsible for the folding, as well as stability related issues. Hence in the algorithm mentioned later this property is taken as the key in preprocessing of the input sequence, i.e. the binary representation where '1' denotes the hydrophobic amino acids and others as '0', as per the hydrophobicity scales proposed by Kyle et. al [9]. Then using the complex plane the folding configurations are formed and their combinations denote various turns [10]. The cumulative sum of the configuration is calculated which gives the direction of each fold. Later the free energy of each folding is calculated using Euclidean distance between the hydrophobic amino acids i.e. all 1s and as per the study the folding having lower free energy value would be stable hence the stable structures could be obtained.

B. *Data*

The data in this case is a protein sequence loaded from protein data bank with pdb id 5CYT, heme protein, using Matlab 7.

Pro=
'XGDVAKGKKTFVQKCAQCHTVENGGKHKVG
PNLWGLFGRKTGQAEGYSYTDANKSKGIVWN
NDTLMEYLENPKKYIPGTKMIFAGIKKKGERQ
DLVAYLKSATS'

C. *Methods*

In brief the steps are as follows:

1) Preprocessing of the input primary protein sequence using the hydrophobicity scale developed by Kyte & Doolittle [9], i.e. developing a vector with hydrophobic amino acids represented by 1 and hydrophilic by 0.
2) Calculating the free energy of this initial sequence
3) Now generating various foldings through iteration, using complex number 'i'.
4) Calculating the free energy for all these foldings.
5) Now further these free energy values could be used to check the stable structures.

D. *Algorithm*

Input – an amino acid sequence, Pro.

Output- an array of free energy of each structure predicted, E.

1) Preprocessing of the input protein sequence
    a) N ← length(Pro)
    b) bin ← Pro
    c) for idx ← 1:N
    d) if Pro(idx)= hydrophobic
    e) then bin(idx)← 1
    f) else bin(idx) ← 0
    g) end
    h) end
2) folding formation
    a) conf ← ones(length(bin)-1,1)
    b) e ← Free_energy(conf)
    c) for k ← 2:length(conf)
    d) f(1:k) ← i
    e) f(k+1:end)←1

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

f) conf ← conf*f
g) F(:,count) ← conf
h) count = count+1
i) end

3) free energy of all the structures in F(m,n)
   a) for j ← 1:n
   b) q ← F(:,j)
   c) p ← Cumulative_sum(q)
   d) E(j) ← Free_Energy(p)
   e) End
4) Algorithm for Cumulative Sum
   Cumulative_sum (a)
      a) for x ← 1 : length(a)
      b) sum ← sum + a(x)
      c) end
5) Algorithm for Free energy
   Free_Energy(a)
      a) a ← a * (bin with only hydrophobic elements)
      b) for x ← 1 : length(a)
      c) d ← abs(a(x) –a(x+1))
      d) sum ← sum + d
      e) end
      f) energy ← sum

IV. Results

The length of the sequence in this case was 104, hence as the algorithm total number of folding created is 103, each column of matrix F (fig. 1) shows a folding. And each row of array E (fig. 2) shows the free energy for each folding. Here the free energy of the unfolded structure is 'e= 45194'.

V. Discussion and futurework

The result from this approach provides the practical aspect of the impact of hydrophobicity on stability, the various outcomes could be used for further research or with some modifications could lead the ultimate solution. With the help of this method the folding could be generated at any structure level, these folding could be used for further research work like in machine learning or neural networks. The free energy calculated could be further used for clustering or classification purposes, thus could enhance the study of the stability factors. In the future work hydrophobicity could be coupled with any other amino acid feature.

REFERENCES

[1] Matthew J. Betts and Robert B. Russell , Amino Acid Properties and Consequences of substitutions, Chap. 14, 'Bioinformatics for Geneticists', 2003.
[2] Cuff JA. Barton G.J. "Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction, PROTEINS: Structure, Function, and Genetics, 1999; 34: 508-19, Available from: httop://binf.gmu.edu/vaisman/csi731/pr99-cuff.pdf.
[3] Zhe Zhang, Lin Wang, Daquan Gao, Jie Zhang, Maxim Zhenirovskyy and Emil Alexov, "Predicting folding free energy changes upon single point mutations". Bioinformatics Advance Access published, Jan. 2012.
[4] Shaolei Teng, Anand K. Srivastava, and Liangjiang Wang, "Biological Features for Sequence-Based Prediction of Protein Stability Changes upon Amino Acid Substitutions". International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, 2009.
[5] H.C. Gasteiger E., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D. and Bairoch A. , The Proteomics Protocols Handbook, Humana Press, 2005.
[6] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," Nucleic Acids Res, vol. 28, Jan 1. 2000, pp. 374.
[7] Lecture 2, Proteins: structure, translation, http://www.sfu.ca/~leroux/class_L02.ppt.
[8] Protein stability, Protein Folding, misfolding – chemistry, http://www.chemistry.gsu.edu/faculty/.../Protein /lecture6_foldingprotein_stability.ppt
[9] 76-456/731 Biophysical Methods- Protein structure component, Lecture 2: Protein interactions leading to folding http://www.chembio.ugo.
[10] Jack Kyte and Russell F. Doolittle,' A simple method for displaying the hydropathic character of a protein ', J. Mol. Biol. 157, 105-132, 1982.
[11] www.mathworks.in/matlabcentral/contest/contests/11/rules.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

AUTHORS PROFILE

R. C. Jain, M.Sc., M. Tech., Ph. D., is a Director of S.A.T.I. (Engg. College) Vidisha (M. P.) India. He has 37 years of teaching experience. He is actively involved in Research with area of interest as Soft Computing, Fuzzy Systems, DIP, Mobile Computing, Data Mining and Adhoc Networks. He has published more than 125 research papers, produced 7 Ph. Ds. and 10 Ph. Ds are under progress.

Geetika S. Pandey obtained her B.E degree in Computer Science and Engineering from University Institute of Technology, B.U, Bhopal in 2006. She obtained Mtech degree in Computer Science from Banasthali Vidyapith, Rajasthan in 2008. She worked as Assistant Professor in Computer Science and Engineering Department in Samrat Ashok Technological Institute, Vidisha (M.P). She is currently pursuing Ph.D. under the supervision of Dr. R.C Jain, Director, SATI, Vidisha. Her research is centered on efficient prognostication and augmentation of protein structure using soft computing techniques.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Appendix

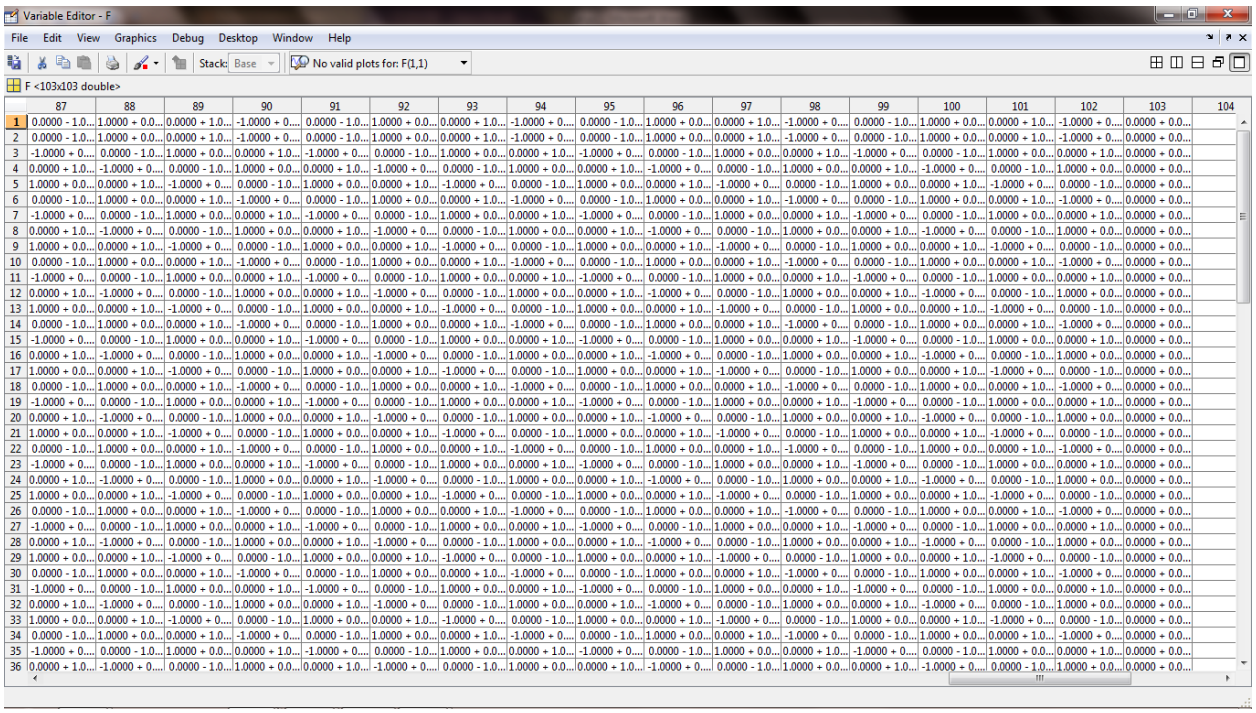


Fig. 1, F(103x103) , various folding of sequence pro.

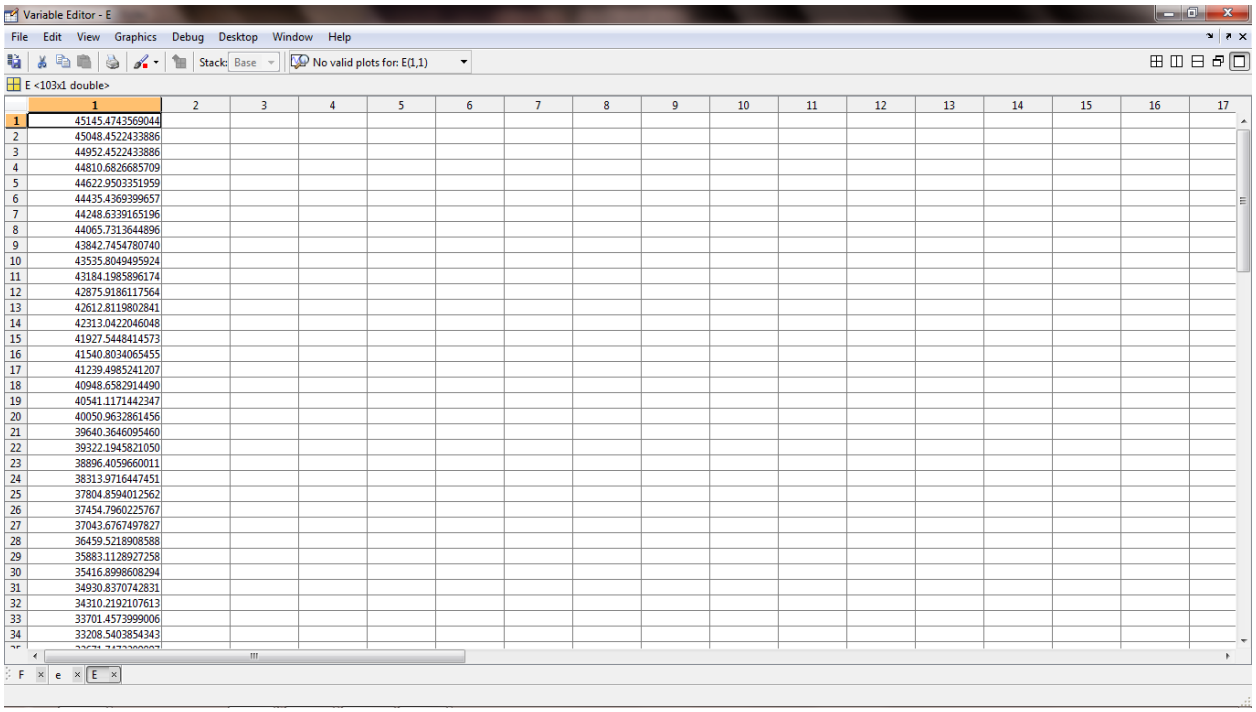


Fig. 2, E(103x1), free energy of each folding.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

SURVEY ON MAC PROTOCOLS USED IN CO-OPERATION NETWORK USING RELAY NODES.

Shalini Sharma
PG-Scholar
DoEC, SSSIST

Mukesh Tiwari
Associate Professor
DoEC, SSSIST

Jaikaran Singh
Associate Professor
DoEC, SSSTTS

***Abstract*— In this paper a survey on a relay based media access scheme has been proposed. It has been observed that any cooperative scheme gives better performance by availability of additional path using the concept of a relay nodes. Relay based schemes more than one relay nodes are selected to improve the performance, so that if one fails the other can be used as a back. Such a co-operative scheme will enhance the performance of the network.**

Keywords: MAC, cooperation, Relay Performance, Newtwork. Scheme.

1. INTRODUCTION

With the advent of new wireless networking applications and wireless industries, our world has dramatically changed from large, low speed, low output devices and applications to high speed, high-performance devices like mobile phones, laptops, navigators, cordless phones. Gaming consoles, etc. and applications. All this is made possible with the use of low-cost and high data rate IEEE 802.11 [20] based ad-hoc networks [10].IEEE 802.11 network support two different medium access control (MAC)[4] mechanisms. One is called distributed co-ordination function (DCF) [15] which helps the mobile nodes to spontaneously form an ad-hoc network. The other is called point coordinated function (PCF) [2] which is mostly used in infrastructure based network.

1.2 Ad-hoc Network

A wireless Ad-hoc network [8] is a decentralized type of network, and it does not rely on a pre-existing infrastructure, such as routers or access points. Each node itself acts as a router to take its decision to route or relay the data packets.

1.3 Co-operative Network [3]

A network in which the neighboring nodes located at the source, and the destination help in the data transmission from the source to the destination. These neighboring nodes may be high data rate nodes and may improve the transmission by reducing the delay and improving the throughput.

1.4 Relay Node/Helper Node

A relay node [4] is an intermediate node which is a neighbour or closer to both the source node and the destination node. And the function of the relay node is to implement co-operation between the source and the destination to forward the data.

2. LITERATURE SURVEY

The thinking about the concept of relays started in 2003 with rPCF (Relay Point Co-ordinated Function) [9]. If the direct link has a low data rate and there exists a relay node such as that links from source to relay and relay to a destination provides better data rate, then transmission can proceed to use the relay node. Further same concept has been applied in Distributed Co-ordination Function (DCF) by introducing rDCF (Relay Distributed Co-ordination Function) [7]. A willing list is maintained by each neighboring node (Nr), and an entry (Ni to Nj) is added if the Nr finds that the transmission rate is improved if done via Nr. This willing list is periodically advertised. If Ni gets this willing list from Nr and finds its entry, then it adds it to its relay table. Another protocol named EMR (Efficient Multirate Relaying MAC) [5] works as a secondary protocol over primary network layer protocol which forms the main route and the main route is converted into multirole route by using EMR For the relay selection the effective throughput is calculated for various combinations of source relay and destination which is mapped to a priority value. S. Zou [1] has described RAMA protocol which also makes use of multihop high rate links, in the network. Once the source gets the invitation triggered by the relay node it adds the relay in its relay list. When a source has data to send to the destination it checks the relay list, and uses it for relaying the data. Another protocol called DAFMAC (Decode and Forward MAC) [2] uses distributed timers for the selection of relays, where each potential relay transmits only after its delay timer expires. If the relay node does not hear ACK after SIFS duration it sends data to the destination. Another timer based relay selection protocol is the CCBF [6] ,which uses a metric called EADV (expected advance). The node having positive value of EADV starts a timer. The node whose timer expires sends a CTS message, in response to the RTS sent by source. CoopMAC (Co-operative MAC) protocol [3], [16] proposed by P. Liu, Z. Tao, S. Narayan uses RSSI(Received Signal Strength Information) for the relay selection , RSSI is also used RM-MAC (Relay multirate MAC) [10]. Here the path via relay node is chosen only if better data rate is achieved .But in

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

UtdMAC (University of Texas Dallas) [12] path via relay node is only kept as a backup in case direct transmission is failed. CODE protocol [13] uses simultaneous transmissions by multiple relays to achieve power gain. This protocol considers both the co-operative transmission using multiple relays along with network coding when the traffic is bidirectional. In RID (Relay with Integrated Data) [14] as well as ARC-MAC (Active Relay Based Co-operative MAC) [12] the high data rate relays which are used to assist transmissions also gain by helping. When the source transmits data to destination via relay, the relay encapsulates its data packet into the source data packet, and this new combined data packet is sent to the destination. In CORELA (Co-operative Relaying Enhanced Link Adaptation) [15] an enhanced link adaptation algorithm has been described. The protocol gives advantage of both the co-operative relaying and link adaptation. By co-operative relaying, reliability is enhanced and by link adaptation, the bandwidth efficiency is improved. In[19] described a co-operative MAC Protocol which makes use of two best relays on the basis of the data rate information sent by them in the RR (Relay Response) frame .The second relay is the backup relay. The two best relays are selected such that total transmission time through the first relay path plus the backup relay path is less than the direct transmission time between the source and destination.

3. Various MAC Techniques

IEEE 802.11 provides physical layer multirate capability. This implies that the data can be transmitted at different rates depending upon the physical channel condition. When the signal to noise (SNR) ratio is high then high data rate is to be used to get the necessary bit error rate (BER)[1]. IEEE 802. 11a supports data rates of 6,9,12 … 54 Mbps. Whereas IEEE 802.11 b supports data rates of 1,2,5.5 and 11 Mbps. So, to utilize these capabilities different kinds of MAC mechanisms are needed. Some schemes use direct links for transmission as explained earlier, for example, ARF (Auto Rate Fallback) [18], RBAR (Receiver Based Auto Rate) [20] and the OAR (Opportunistic Auto Rate) [19]. However, they only use the direct link between the sender and the receiver. If the direct link is of poor channel quality, then they do not give higher performance in terms of throughput and delay. Hence co-operative communication i.e. communication using relays or helper nodes, was introduced to get a higher throughput and thus improve the performance of the network.

In the following sections, various co-operative MAC schemes have been explained. Some of these schemes use a single relay, and some of them use two relays. Relay may be used as a backup transmission path, or it may be used as a better path for the transmission. A detailed explanation of the different co-operative scheme is given below.

3.1 AR-CMAC (Active Relay Based Co-operative MAC) :

3.2 DAFMAC (Decode and Forward MAC) :

Decode and Forward MAC (DAFMAC) [11] is efficient co-operative diversity partner selection algorithms for IEEE 802.11. Most of the protocols assumes the availability of relays. But this protocol helps in efficient selection of relay for co-operation. Each node is associated with a timer t. The relay selection process uses distributed timers where each potential relay transmits only after its delay timer expires. The delay t allocated to each node is inversely proportional to the Received Signal Strength (RSS) .Best node has high RSS and less delay. If the relay node does not hear ACK after SIFS duration, it determines its t. If it hears any helper node co-operating within time interval T = SIFS+t, it goes to idle state. Else, the relay sends data to the destination. After hearing ACK from the Destination, the relay forwards ACK to the Source. Hence, the protocol improves transmission reliability for streaming media.

3.3 CCBF (Co-operative contention-based forwarding):

CCBF [17] is an improvement over the existing co-operative protocols known as CBF (cluster-based forwarding) [2] and CRL (Co-operation using relay and leapfrog) [8] by using CCBF, the scope of co-operation is extended and the forwarding opportunities provided by the available neighboring nodes is fully utilized. Protocol has two contention phases. First is the next hop selection, which is RTS/CTS based and the other is the co-operative forwarding. The protocol follows the usual IEEE 802.11 DCF protocol. In the contention process, a metric called EADV (expected advance) is calculated. The node having a positive value of EADV starts a timer. The node whose timer expires sends a CTS message. Overhearing the CTS other nodes to cancel their timers. On receiving the CTS, the sender sends the data to the next hop receiver. The protocol introduces the concept of the co-operative area within which the nodes contend to become the forwarders, after the sender transmits the data packet to the next hop receiver .Again the forwarder is selected on the basis of a metric. Results show that CCBF has improved latency, energy efficiency and improved routing performance.

3.4 2rcMAC (Two Relay Co-operative MAC):

The 2 relay co-operative MAC Protocol [9] makes use of two best relays or helper nodes, which would enhance throughput and reliability in data transmission between the source and the destination. These nodes are chosen from a set of relay nodes on the basis of the data rate information sent by them in the RR (Relay Response) frame. A relay node based on the source to relay and relay to destination estimation chooses the suitable rate slot and sends a single-bit feedback in a randomly picked bit interval location of the RR frame. The relay nodes after hearing the RTS and CTS frames respond to the source with their rate information, so that the source may select the two best relays according to its transmission requirements. The two best relays are selected such that total transmission time through the first relay path plus the backup relay path is less than the direct transmission time between the source and destination. Having a backup path in case of first relay path failure improves the reliability of the overall system. This protocol is explained further in the later sections as it forms the base of our research work.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

3.5 RM-MAC (Relay multirate MAC):
This protocol described [6] is a combination of two protocols, one named MASA[5] and the other is RBAR[16].A modified version of RBAR is used. This protocol also employs helper nodes to enhance the throughput of the network. Unlike RBAR, the rate information is not received in the CTS message; rather a neighbour node table is used to find the corresponding data rate. The protocol calls the helper nodes as assistant nodes, which record the signal strength of the packets sent from source to destination. The ratio of the two signal strengths is called back off value of the assistant node. When the data packet sent by the source is not correctly received by the destination node, then the assistant node closest to the source nodes and having least ratio is chosen to retransmits the data packet. After the back off value, the assistant node sends SACK packet to the source and SDATA to the destination. The destination on correctly receiving DATA replies to the assistant node with ACK message. If the DATA is not delivered then after fixed number of retrials a failure record is marked in the neighbour table.

3.5 CORELA (Co-operative Relaying Enhanced Link Adaptation):
An enhanced link adaptation algorithm called CORELA has been described in [10]. The protocol gives an advantage of both the co-operative relaying and link adaptation. By co-operative relaying, reliability is improved and by link adaptation, the bandwidth efficiency is improved. This protocol is composed of two modules. One is the co-operative relay module, and the other is the link adaptation module. The two modules are independent of each other, so the changes in one do not affect the other. For the co-operative relay module, SIDF (Selection incremental decodes and forward) scheme, which is an improvement of a DF (Decode and Forward) scheme is used. And in the link adaptation module, a data rate is chosen, which is suitable for relay destination link condition. A number of counts such as transmission, success as well as failure count and relay transmission success and failure counts are maintained. These parameters are used to decide the switching of data rates, hence different transmission rates may be chosen in the direct channel and relay channel maximal channel efficiency.

4. CONCLUSION

In this work literature survey has been done related with MAC protocols. It has been observed that with relay node, performance improves. Wireless network is a compulsion of today life and due to dependency on it, in the journal every time it is tried to increases options for the network to enhance its working.

References

1. A. S. Ibrahim, A. K. Sadek, W. Su, and K. J. R. Liu, "Cooperative communications with Relay selection: when to cooperate and whom to cooperate with?"IEEE Transactions on Wireless Communications, vol 7 July 2008.

2. J. S. Pathmasuntharam, A. Das, and K. Gupta, "Effi cient multi-rate relaying (EMR) MAC protocol for ad hoc networks," in Proc. of IEEE ICC'05, Seoul, Korea, July 2005.

3. J.M. Bahi, M. Eskandar and A. Mostefaoui, "A Robust Cooperation Approach For Mobile Multimedia Ad-hoc Networks," IEEE 2008.

4. J.N. Laneman, G. W. Womell, and D. N. C. Tse, "An efficient protocol for realizing cooperative diversity in wireless networks," in Proc. IEEE International Symposium on Information Theory (ISIT), June 2001.

5. J.P Hubaux, "Stimulating Cooperation in Self-Organizing Mobile Ad Hoc Network," Technical Report No. DSC/2001/046, August 2001.

6. Kaleshi, D., Zhong Fan, "CORELA: A cooperative relaying enhanced link adaptation algorithm for IEEE 802.11 WLANs" 8th International Symposium on Wireless Communication Systems 2011.

7. K. T. Wan, H. Zhu, and J. Andrain, "CODE: Cooper ative Medium Access for Multirate Wireless Ad Hoc Network," in Proc. IEEE SECON, 2007.

8. L. Blazevic, L. Buttyan, S. Capkun, S. Giordano, J. P. Hubaux, and J. Y.Le Boudec, "Self-organization in mobile ad-hoc networks: The approach of terminads," IEEE Commun. Mag., vol. 39, no. 6, pp. 166–174, June 2001.

9. L. Buttyan and J.P. Hubaux, "Enforcing Service Availability in Mobile Ad-Hoc WANs", Proc. of Workshop on Mobile Ad-hoc networking and Computing (MobiHOC), Boston, USA, August 2000.

10. L. Zhang and Y. Shu, "Throughput and Fairness Improvement in 802.llb Multi-rate WLANs," in Proc. of IEEE International Symposium on Personal, Indoor Mobile Radio Communications, pp. 1946-1950, Sept. 2005.

11. Li, D. Kaleshi, Zhong Fan, "A cooperative relaying enhanced link adaptation algorithm for IEEE 802.11 WLANs" 8th International Symposium on Wireless Communication Systems (ISWCS), 2011.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

12. Long Cheng, Jiannong Cao, Canfeng Chen, Hongyang Chen, Jian Ma, Joanna Izabela Siebert, “Cooperative Contention-Based Forwarding for Wireless Sensor Networks." IWCMC ’10, July 2010, Caen, France.

13. M. Jakobsson, J.P. Hubaux, and L. Buttyan, “A Micro- Payment Scheme Encouraging Collaboration in Multi-Hop Cellular Networks." Proceedings of Financial Crypto 2003.

14. Murad Khalid, Yufeng Wang, In ho Ra and Ravi Shankar, “Two Relay Based Co-operative MAC Protocol for Wireless Ad-hoc Network." IEEE Transactions on Vehicular Technology Vol 60, No. 7 September 2011.

15. N. Agarwal, D. ChanneGowda, L. Kannan, M. Tacca, and A. Fumagalli, “IEEE 802.11b cooperative protocols: A performance study,” in Proc. NETWORKING, vol. 4479, LNCS, 2007, pp. 415–426.

16. Nadia Qasim, Fatin Said, Hamid Aghvami, “Mobile ad hoc networking protocols’ evaluation through simulation for quality of service." IAENG International Journal of Computer Science, 36:1, IJCS_36_1_10.

17. “Network Simulator NS-2” [Online]. Available: http://www.isi.edu/nsnam/ns.

18. P. Caballero-Gil, J. Molina-Gil, C. Hernandez-Goya and C. Caballero-Gil, “Stimulating Cooperation in Self-organized Vehicular Network." Proceedings of the 15th Asia-Pacific Conference on Communications (APCC 2009) -082.

19. P. Coronel, R. Doss, and W. Schott, “Geographic routing with cooperative relaying and leapfrogging in wireless sensor networks." In Proc. IEEE GLOBECOM, 2007.

20. P. Liu, Z. Tao, S. Narayan, “CoopMAC A Co-operative MAC for Wireless LANs." IEEE J. Sel. Areas Communication vol. 25 Feb 2007.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IJCSIS AUTHORS' & REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Mrs Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Mr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Mr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Mr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Mr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Mr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Mr. P. Vasant, University Technology Petronas, Malaysia
Mr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Mr. Praveen Ranjan Srivastava, BITS PILANI, India
Mr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Mr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Mr. Tirthankar Gayen, IIT Kharagpur, India
Ms. Huei-Ru Tseng, National Chiao Tung University, Taiwan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2013
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/